

## An interactive bi-objective optimisation process to guide the design of electric vehicle warning sounds

Tom Souaille<sup>1,2</sup>, Jean-François Petiot<sup>1</sup>, Nicolas Misdariis<sup>2</sup> and Mathieu Lagrange<sup>1</sup>

<sup>1</sup>Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

<sup>2</sup>STMS Ircam-CNRS-SU, Paris, France

### Abstract

Electric vehicles (EVs) are very quiet at low speed, which can be hazardous for pedestrians, especially visually impaired people. It is now mandatory (since mid-2019 in Europe) to add external warning sounds, but poor sound design can lead to noise pollution, and consequently annoyance. Moreover, it is possible that EVs are not sufficiently detectable in urban areas because of the masking effect from the background noise. In this paper, we propose a method for the design of warning sounds that takes into account both detectability and unpleasantness. The method implements a multiobjective interactive genetic algorithm (IGA) for the optimisation of the characteristics of synthesised sounds. An experiment is proposed to a first panel of participants in order to define a set of Pareto efficient sounds. At the individual level, sounds obtained with the IGA are compared to different sound design proposals. Results show that the quality of the sounds designed by the IGA method is comparable to those provided by a sound designer. From the sounds of the Pareto set, a design recommendation method based on the probability distributions of the sounds' characteristics is proposed. An external validation with a second panel of participants shows that these recommended sounds constitute relevant trade-offs when compared to other design proposals.

**Key words:** interactive genetic algorithm, subjective evaluation, electric vehicle, sound design, multi-objective optimization

Received 06 December 2021

Revised 01 August 2022

Accepted 01 August 2022

Corresponding author

T. Souaille

tom.souaille@ls2n.fr

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Des. Sci., vol. 8, e26

[journals.cambridge.org/dsj](https://journals.cambridge.org/dsj)

DOI: [10.1017/dsj.2022.18](https://doi.org/10.1017/dsj.2022.18)

the **Design Society**  
a worldwide community

 **CAMBRIDGE**  
UNIVERSITY PRESS

### 1. Introduction

The increasing use of Hybrid and Electric Vehicles (EVs) in recent years, also called quiet vehicles (QVs), has led to safety concerns for pedestrians. Below 40 km/hour, the noise emitted by those vehicles is lower than for internal combustion engine vehicles (ICEVs; Japan Automotive Standards Internationalization Center 2009). In urban environments in particular, this can make it more difficult to detect an approaching vehicle. Visually impaired people are particularly affected, as they rely mostly on auditory cues to assess the presence of vehicles (Konet *et al.* 2011; Parizet, Ellermeier & Robart 2014). Because of this, legislation now exists in several countries, requiring QVs to be equipped with a warning sound generation device (acoustic vehicle alert system), as well as specifications regarding the sound that

should be emitted (Lee *et al.* 2017). Nevertheless, regulations regarding the sound of electric cars may evolve and their sound design remains widely open. Several studies came up with recommendations regarding the nature of such a sound, based on detection time measurements (Misdariis, Gruson & Susini 2013; Robart & Parizet 2013; Parizet *et al.* 2014; Poveda-Martínez *et al.* 2017). However, these recommendations should also take into account potential noise pollution that could negatively affect the experience of pedestrians, cyclists and other drivers (Petiot, Kristensen & Maier 2013).

It is indeed clear that QV sounds may be masked by the background noise of the environment, making them hard to detect. A naive solution which consists of a simple increase of the sound level to reduce the masking effect may have dramatic consequences on sound pollution. Thus, there is a conflict between detectability and annoyance in the perception and design of QV sounds. Different studies addressed this problem (Parizet *et al.* 2014; Singh, Payne & Jennings 2014; Lee *et al.* 2017; Steinbach & Altinsoy 2019). Between 2010 and 2014, the European project eVADER showed that it is possible to design a QV warning sound that is both easily detectable and of a low-amplitude level (Parizet *et al.* 2014). However, these approaches do not allow for an extensive sound space exploration, as the sounds tested are chosen from a fixed, small corpus of artificial sound stimuli.

Furthermore, the design of QV sounds is also crucial from a branding point of view. Though being partially limited by the legislation, a manufacturer should be able to explore a wide range of possibilities, in order to design a sound that will distinguish them from their competitors. In that respect, anticipating customer preferences and predicting the perceived quality is important (Swart, Bekker & Bienert 2018). QV is a rather new product, and its collective representation in terms of sound identity has yet to be defined. This could also be an opportunity to make bold choices about the futuristic image of the vehicle, if that is considered a good selling point. One may also desire a QV sound that is as similar as possible to an ICEV.

The design of QV warning sounds is clearly a multiobjective design problem that is closely related to human auditory perception. The challenge for the sound designer is to understand all the facets of the problem and to translate them into relevant acoustic attributes. In addition to the expertise of a designer (and his/her innovativeness! [Engler 2016]), listening tests are required to understand the complex relationships between acoustic parameters and perceptual dimensions (Edworthy, Loxley & Dennis 1991). Therefore, to assist designers in their design decisions and to confirm their proposals, an active research field in product design considers the analysis of end users' perceptions or preference, to extract useful information to make design decisions (Orsborn, Cagan & Boatwright 2009; Bi *et al.* 2017).

The purpose of this paper is to define a tool to guide the design of product sounds. The proposal is based on an interactive genetic algorithm (IGA) because of the potential of such algorithms to interactively take into account feedback from the listeners, and their ability to consider conflicting objectives.

In Petiot, Legeay & Lagrange (2019), a study using an IGA for the design of QV sounds was presented. After a definition of the experimental protocol for the listening tests and the sound synthesis technique, a mono-objective optimisation using IGA was implemented, to explore the trade-off between detectability and unpleasantness, which were aggregated as a weighted sum. The results showed that the sound solutions produced by the IGA were efficient when compared to

proposals of a designer. In a following work (Souaille *et al.* 2021), we showed that interesting recommendations can be extracted from the analysis of Pareto efficient QV sounds, provided by a multiobjective IGA experiment (optimisation of detectability and unpleasantness).

This paper is a continuation of these two studies. It uses the same experimental protocol for the assessment of the detectability and the unpleasantness of sounds, but validation experiments are added to demonstrate the efficiency of the proposals.

The general aim of the study is to propose a method based on listening tests and an IGA to assist the design of alert sounds for QVs. A multiobjective IGA is implemented with a panel of participants, in order to produce a set of Pareto efficient solutions. A method is next proposed to define design recommendations, from the analysis of the Pareto set. These recommendations are then compared to other design proposals (provided by a designer, or randomly defined). Two research questions are considered:

- (i) Q1: At an individual level (for each participant), how do QV sounds designed with the IGA compare to other design proposals, in terms of unpleasantness and detectability?
- (ii) Q2: How do the design recommendations obtained with the proposed method (by gathering individual data) compare to other design proposals, in terms of unpleasantness and detectability, with an external panel of listeners?

Three experiments are carried out to answer these questions. In Experiment 1, participants are invited to use the IGA paradigm in order to define individual Pareto optimal QV sounds, from which a single 'best' sound is selected, for each participant. In Experiment 2, the same participants have to assess their 'best' sound, together with other proposals, in order to compare their performances. In Experiment 3, a second panel of participants is invited to assess the recommended designs, together with other design proposals, in order to compare their performances.

The two novel contributions of the work are:

- (i) the use of an IGA applied to a multiobjective sound design problem;
- (ii) a solution recommendation method, based on the analysis of the solutions obtained by a group of subjects, who used an IGA applied to a multiobjective problem.

For the design of a QV sound, these tools aim at helping a sound designer, by the proposition of efficient sound examples, or by the tuning of synthesis parameters. The proposed method could be transposed to other sound design problems with multiple conflicting objectives. For example, a self-driving car manufacturer could want to design sounds that continuously inform the passengers of surrounding traffic conditions and driving behaviour, while not being too intrusive or annoying (Misdariis, Cera & Rodriguez 2019; Fagerlönner, Larsson & Maculewicz 2020). More generally, the recommendation method could be applied to any multiobjective design problem dealing with the interactive optimisation of perceptual quantities.

The paper is organised as follows. A short background on the integration of user perceptions in product design is presented in Section 2, together with a description of the multiobjective IGA implementation used in this study. Section 3 presents the sound synthesis method implemented, the listening test

scenario and the interfaces. Sections 4–6 are dedicated to the description of the three experiments, with the methods implemented and the analysis of the results. Section 7 proposes a summary of the main results. Section 8 presents and discusses the results of the experiments, and concludes on the hypotheses under study. In Section 9, conclusions are drawn on the main contributions of this paper and recommendations for the design of sounds are made.

## 2. Background

### 2.1. Integration of user perceptions

In order to include users' perception in the design process, two categories of methods can be considered.

The first category of methods to tackle this problem is based on the modelling of users' perceptions or preferences according to a given set of parameterised products (modelling of perceptual data). These methods use the design of experiments (DOE) theory and assume an explicit model (generally linear without interaction) between the perceptual dimensions and the design variables. Various statistical procedures and experiment designs exist to estimate the coefficients of the model. The Japanese Kansei engineering, a well-known design method to account for users' feelings and perceptions (Nagamachi 1995), belongs to this category. For instance, it is considered in sound design to explore how human feelings and emotions can be evoked by a sound's physical properties. For warning sounds, many studies proposed an experimental approach with listening tests to understand human perceptions for sound design (Ibrahim, Yiap & Andrias 2018). In Marshall, Lee & Austria (2007), different listening scenarios are proposed to study two objectives, namely annoyance and urgency, with a fixed DOE. Knowing the effect of sound parameters on the perceived urgency is important to give precise recommendations (Edworthy, Loxley & Dennis 1991). Another approach consists in testing how classical psychoacoustic metrics (loudness, roughness, fluctuation strength etc.) can explain perceptual assessments (Lee *et al.* 2017).

To study and understand human reaction to sounds, experiments generally use a parameterised sound synthesis method and model-based DOE (e.g., D-optimal DOE). The limitation of such an approach is that a model between the acoustic parameters and the perceptual dimension must be assumed in advance, given that the exact form of the model is generally unknown. To reduce the complexity of the models, the sounds proposed in the listening tests are generally simple stimuli that are not representative of the complexity of real design solutions and fail to be relevant for multiple objectives. Furthermore, a systematic exploration of a large sound space can be a tedious task – or even infeasible – and user listening fatigue has to be taken into account.

The second category of methods for the analysis of users' perceptions is not model-based and uses human–computer interactions. These methods are model-free in content (contrary to DOE, there is no model of the behaviour of the respondent), but model-driven for the solution search. In this case, an algorithm gradually refines the propositions made to the users. In interactive evolutionary computation (IEC) algorithms, for example, the user plays the role of the evaluator in an evolutionary process (Takagi 2001). In IEC, the user assesses the fitness of a

design population (which is the adaptation of the population to the problem), by rating the proposed designs or choosing the best ones, for example. IEC has been applied to many domains (music, writing, education, food industry etc.) involving different sensory modalities.

Particular cases of IEC are IGAs, where genetic operators, such as recombination, crossover and mutation, are used to modify the designs throughout the optimisation process. This method has been used, for example, to capture the aesthetic intention of participants for the design of cartoons (Gu, Tang & Frazer 2006). IGAs have also been tested in our previous studies for the design of car dashboards (Poirson *et al.* 2013), which have confirmed their relevance for extracting design trends and obtaining a final product solution that optimises a determined semantic dimension. IGAs have the advantage of not needing restrictive assumptions regarding the perceptual model of the participant. Interaction effects are in fact implicitly integrated in the course of the model-driven search in the solution space. IGAs have been used in sound design for musical compositions (Biles 2007) or to design sign sounds, with the purpose of communicating a message through a melody (Miki *et al.* 2006). Subtle perceptual phenomena can be taken into account for the optimisation of products involving sensory constraints, such as the tuning of cochlear implants (Wakefield *et al.* 2005).

Nevertheless, the implementation of IGAs to support sounds design remains rare. The first reason is that the design of products' sounds is a relatively recent discipline that does not have as well-established guidelines as in visual design, where products can be described by illustrations or advanced 3D models (Eppinger & Ulrich 2016). Conversely, sounds are immaterial and embedded in time, which makes the definition of prototypes difficult. 'Sound sketches', using, for example, voices and gesture (Delle Monache *et al.* 2018), can be proposed to rough out a design proposal, but their use still remains limited. The second reason is that the design of sounds is critically dependent on the process of listening. According to Özcan & Van Egmond (2012), products' sounds remain mainly based on the subjective experience of the designer because perceptual factors are complex to investigate. Several perceptual dimensions, potentially conflicting, may be elicited when listening to a sound. And even if generic tools based on a shared vocabulary and sound examples can be proposed (Carron *et al.* 2017), their adaptation to a particular project remains difficult.

Comparisons of IGA and classical DOE are rare in the engineering design literature. DOE allows for the definition of optimal products, the estimate of effects of factors on the response and the generation of predictions in the design space (with the model), while IGA is only oriented towards the search for the optimal product. A comparison of IGA and fractional factorial design for the design of warning sounds is presented in Petiot *et al.* (2020). Results show that the IGA method can be an interesting alternative to classical DOE to help the design of sounds. Another comparative study concerns the design of the shape of a bottle using Conjoint Analysis, where the authors conclude on the superiority of interactive evolutionary algorithms on fractional DOE to elicit the optimum product, but without explanations for the reasons behind this superiority (Teichert & Shehu 2007).

## 2.2. Multiobjective IGA

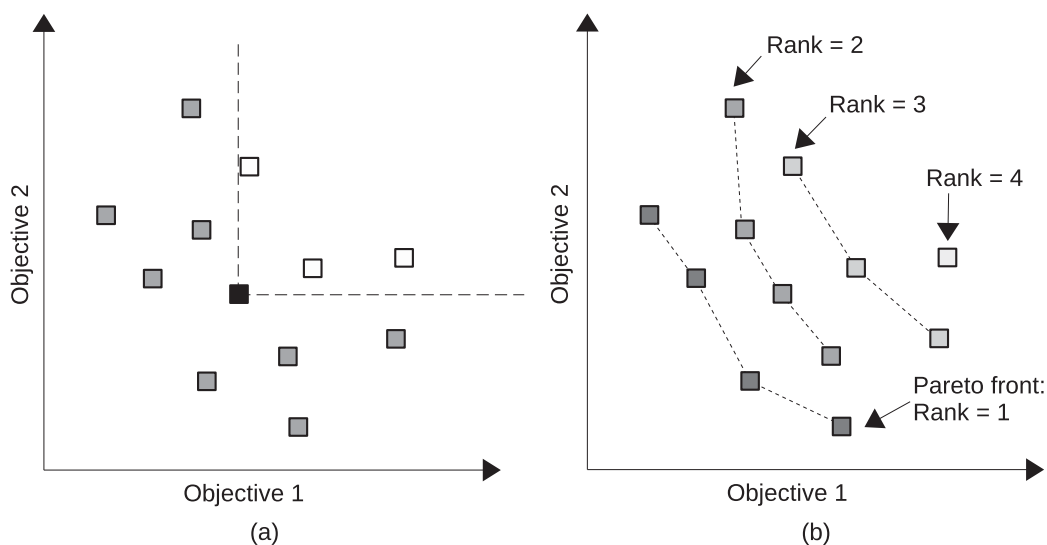
Genetic algorithms take inspiration from some knowledge of the evolutionary process of living organisms to solve optimisation problems (Goldberg 1989). Potential solutions to a given problem are coded as strings of numbers called chromosomes. A chromosome is composed of several sections, called genes, each coding the values of different parameters of the corresponding solution. Solutions are evaluated in groups, called generations, usually using a fitness function. At each generation, a mating pool is created based on previous solutions' performances with respect to the problem's objectives. So-called genetic operators are applied to the solutions within that mating pool to create the next generation of solutions that will be evaluated. This process is repeated until convergence is achieved or the maximum number of generations is reached.

In IGAs, the evaluation of the solutions (fitness) is done by a human, so the fitness function does not have a mathematical expression. With this approach, it is possible to find solutions to problems involving a semantic dimension (such as the notion of 'unpleasantness' in this study, or the 'clearance' of a car dashboard in Poirson *et al.* 2013), or perceptual dimensions. IGA can be used to assist the design of hearing aids (Durant *et al.* 2004) or to optimise the affordances of steering wheels (Mata *et al.* 2018). In this type of procedure, user fatigue is the main limitation to the maximum number of possible evaluations. This limits both the number of solutions per generation as well as the number of generations, which has an impact on the convergence of the algorithm. What is more, for time-varying solutions such as sounds, only one solution can be evaluated at a time.

The optimisation problem addressed in this paper is bi-objective, which means that there are potentially several equally satisfying solutions (Pareto efficient). To address this constraint, the proposed method considers an adaptation of the NSGA-II algorithm (Nondominated Sorting Genetic Algorithm II; Deb *et al.* 2002), which aims at finding an approximation of the optimal Pareto front of the potential solutions. A Pareto front comprises the solutions that are nondominated in the Pareto sense. A solution is said to Pareto-dominate another one if it is better or equal to this solution for all objectives and if there is at least one objective where it is strictly better. An example of this is shown in Figure 1, where solutions are represented according to two objectives that must be minimised. The group of nondominated solutions is called the Pareto front. By ignoring the solutions of the Pareto front, it is possible to find the second front (Rank 2) of nondominated solutions, and so on. Figure 1b shows an example of the Pareto front (Rank 1) and other successive ranks.

In the NSGA-II algorithm, the solutions are compared based on the so-called crowded-comparison operator. A solution is considered better than another one if it has a lower (closer to 1) nondomination rank. The nondomination rank of a solution corresponds to the nondominated front it belongs to. Within a nondominated front, the solutions are ranked based on their distances to other solutions of the same front in the objective space. Solutions that are further away from other solutions are considered better. This aims at ultimately obtaining solutions that are evenly spread along the optimal Pareto front.

In this elitist algorithm, a register of the best solutions evaluated is updated after each generation and is used to create the following one. This register has the same size  $N$  as the number of solutions in a generation's population. After each evaluation of a generation of  $N$  solutions, the register is updated by keeping the



**Figure 1.** Example of Pareto domination for a bi-objective problem where both objectives have to be minimised. In (a), the black square Pareto dominates the white squares. In (b), the nondomination ranks are shown.

best  $N$  solutions amongst the union of the  $N$  solutions in the current generation and the  $N$  solutions that are currently in the register. At the first generation, the register being empty, it is initialised with all  $N$  solutions within the population.

### 2.3. Implementation of the IGA

Our implementation uses the NSGA-II algorithm (Deb *et al.* 2002). The procedure is implemented as follows: the first generation of  $N$  solutions is generated using a Latin Hypercube Sample. After evaluation by the user, the best-solution register is initialised with all  $N$  solutions of this first generation. The next generation is created by randomly applying one of the following genetic operators to each solution within the register, which constitutes the mating pool:

- (i) Mutation: the solution is replicated to the next generation, with one gene value randomly changed.
- (ii) Crossover: another solution is selected within the register through a binary tournament based on the crowded-comparison operator. The chromosomes representing the two solutions are combined, in order to create a new one. This is done by randomly selecting a chromosome location, splitting each chromosome into two parts around that location and connecting the first part of one solution with the second part of the other one. The order in which the solutions are combined is chosen randomly.
- (iii) Selection: the solution is replicated in the next generation without any modification.

The probability for each operator to be applied is defined by a rate  $m_r$ ,  $c_r$  and  $s_r$ , corresponding to mutation, crossover and selection, respectively. These rates are normalised so that they add up to 1, that is,  $m_r + c_r + s_r = 1$ .



To select which operator to apply, a number  $r$  is randomly generated between 0 and 1:

- (i) If  $0 \leq r \leq m_r$ , a mutation is applied.
- (ii) If  $m_r < r \leq m_r + c_r$ , a crossover with another solution is applied.
- (iii) If  $m_r + c_r < r \leq 1$ , a selection is applied.

This process is repeated at each generation. The size of the register is kept constant, only containing the  $N$  best solutions.

The same solution (sound) can thus be evaluated several times during the optimisation process, and some variability is expected with regard to the evaluation of the two objectives by the subject. Thus, there is a risk that the best-solution register would contain several times the same sound, with different objective values. This would reduce diversity in the mating pool, which could lead to a premature convergence. To prevent this, each solution in the register is only present once, associated with the latest evaluation that the subject provided for this sound.

### 3. Sound synthesis and experimental protocol

#### 3.1. QV sound synthesis

The QV sounds are synthesised using the additive synthesis technique (Roads 1996). By considering an analysis of current sounds of different carmakers (Misdariis *et al.* 2012) and personal propositions (Petiot *et al.* 2013), the generation of different but plausible sounds for an electric car includes two types of sounds: (1) a ‘harmonic’ sound (discrete distribution of energy with respect to frequency – the frequencies being multiple of a fundamental) and (2) a noise sound (continuous distribution of energy with respect to frequency). The ‘harmonic’ sound is made of two components: a motor sound ( $C_1$ ), mimicking a combustion engine, and a major chord ( $C_2$ ), giving a tonal component. The noise sound is also made of two independent noise bands ( $C_3$  and  $C_4$ ). More formally:

- (i)  $C_1$ , the engine-like sound, is simulated with a weighted sum of harmonics corresponding to a four-cylinder internal combustion engine (ICE; with RPM the rotation speed of the motor (tr/mn), the fundamental frequency is given by  $f_1 = \frac{RPM}{60}$ ).

$$C_1(t) = \sum_{i=1}^6 a_i \sin(2\pi f_i t). \tag{1}$$

Six subharmonics or harmonics are considered  $f_i \in \{0.5f_1, f_1, 1.5f_1, 2f_1, 4f_1, 6f_1\}$  with the corresponding amplitudes  $a_i \in \{0.2, 0.4, 0.5, 0.2, 0.4, 0.6\}$  (see Desoeuvre *et al.* 2008 for more information on the acoustics of thermal engines).

- (ii)  $C_2$ , the major chord, is made of three harmonic (periodic) notes (root note [fundamental frequency  $f_2$ ], major third [ $\frac{5}{4}f_2$ ] and fifth [ $\frac{3}{2}f_2$ ]). Each note is composed of six harmonics (e.g., for the root note, harmonics  $f_{2i} \in \{f_2, 2f_2, 3f_2, 4f_2, 5f_2, 6f_2\}$  with the corresponding amplitudes  $a_{2i} \in \{1, 0.4, 0.4, 0.1, 0.1, 0.1\}$ ),



$$C_2(t) = \sum_{i=1}^6 a_{2i} \left( \sin(2\pi f_{2i} t) + \sin\left(2\pi \frac{5}{4} f_{2i} t\right) + \sin\left(2\pi \frac{3}{2} f_{2i} t\right) \right). \quad (2)$$

(iii)  $C_3$ , the first noise component, is built as a sum of 150 sines with random phase and frequency,

$$C_3(t) = \sum_{i=1}^{150} \sin(2\pi f_{3i} t + \phi_{3i}), \quad (3)$$

with  $f_{3i} \in [0, 2f_3]$  and  $\phi_{3i} \in [0, 2\pi]$ .

(iv)  $C_4$ , the second noise component, is identical to  $C_3$ , but with another frequency range. The frequency  $f_4$  is chosen so that  $C_4$  has a wider frequency range than  $C_3$ ,

$$C_4(t) = \sum_{i=1}^{150} \sin(2\pi f_{4i} t + \phi_{4i}), \quad (4)$$

with  $f_{4i} \in [0, 2f_4]$  and  $\phi_{4i} \in [0, 2\pi]$ .

The resulting sound  $s(t)$  is the weighted sum of these four components, to which amplitude modulation is applied, with modulation index  $m$  and modulation frequency  $f_m$ :

$$s(t) = (1 + m \cdot \sin(2\pi f_m t)) \cdot (a_1 \cdot C_1(t) + a_2 \cdot C_2(t) + a_3 \cdot C_3(t) + a_4 \cdot C_4(t)). \quad (5)$$

The four amplitude coefficients  $a_i$  are normalised so that

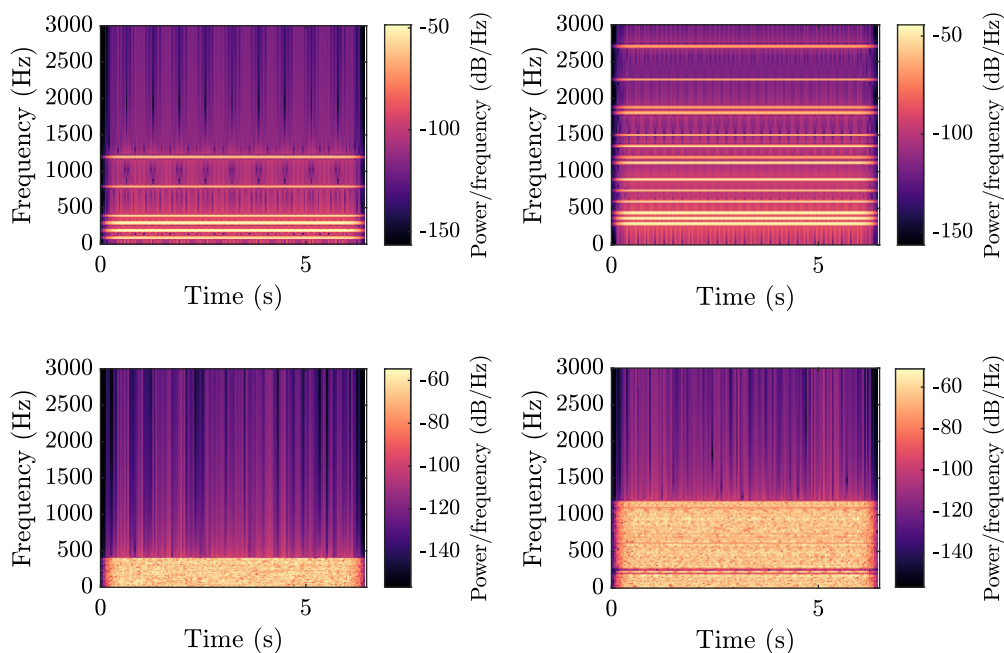
$$\sqrt{\sum_{i=1}^4 a_i^2} = 1, \quad (6)$$

and the ratio between  $a_3$  and  $a_4$  is chosen equal to 0.5.

Figure 2 illustrates each component by showing spectrograms (representation of the evolution of the spectrum of the signal with respect to time) of a sample, for a vehicle driving at constant speed. The horizontal stripes in the top spectrograms denote a harmonic structure made of pure sinusoids, whereas the wide horizontal bands in the bottom plots indicate noise. There is no variation over the horizontal axis, because the sounds are stationary, apart from a short fade in and fade out. Example of sounds can be listened to at <https://mathieulagrange.github.io/souaille2021interactive/>.

The definition of the structure of the synthesised sound and the choice of the variables is the result of many tests, innovative proposals and sound engineering experience of the authors. A complete justification is out of the scope of this paper, the contribution being centred on the optimisation of a given parameterised synthesiser. It is out of the scope of this paper to describe all the parameters of the synthesiser (there are more than 70 independent parameters to define a sound). We can mention that all the frequencies and amplitudes of the components are adjustable, to create credible and original sounds. The synthesiser is controlled by the speed of the car. In this study, this parameter is irrelevant given that the listening tests are made with a constant speed of the car. Readers interested by the design of synthesised QV sounds may consult Pedersen *et al.* (2011) or Petiot *et al.* (2013) for more information on the evolution of the sound according to speed.

Among the different synthesis parameters of the sounds, it is necessary to define the optimisation variables of the problem, that is, the variables that are manipulated by the IGA and coded in the genome (design space of the genetic



**Figure 2.** Spectrograms of the four synthesiser components, without amplitude modulation. Top-left:  $C_1$  motor sound, with  $f_1 = 200$  Hz. Top-right:  $C_2$  chord sound, with  $f_2 = 300$  Hz. Bottom-left:  $C_3$  first noise, with  $f_3 = 200$  Hz. Bottom-right:  $C_4$  second noise, with  $f_4 = 600$  Hz. The brighter the colour, the more power there is in the corresponding time–frequency bin.

code). After several experiments, the following six factors (A, B, C, D, E and F), and their corresponding levels (A1 for Level 1 of Factor A), are chosen to get a large diversity of sounds (see Table 1). These factors control the frequency content and the amplitudes of the synthesiser components, as well as the amplitude and frequency of the amplitude modulation. The choice of the factors is motivated by previous findings showing that harmonic complexity and amplitude modulation are related to detectability (Parizet *et al.* 2014).

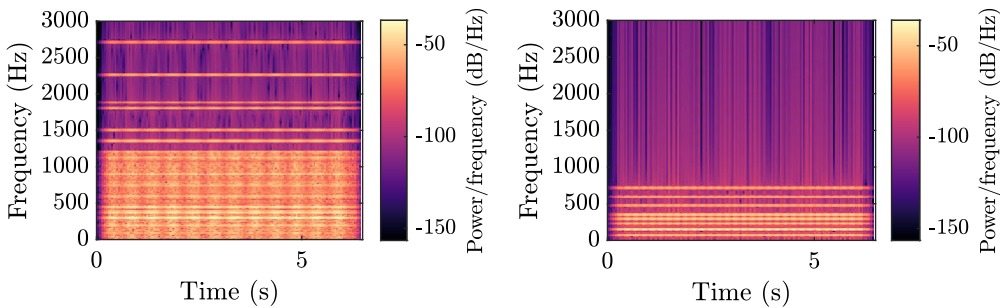
Furthermore, it has been shown that the preference for an EV sound can be related to how similar it is to an ICE car sound (Petiot *et al.* 2013). To take that into account, the synthesis method is able to create sounds that resemble more or less an ICE car sound, depending on the value of Factor A. The setting of the levels of the factors required many adjustments (not reported here) to obtain audible differences between sounds, but with still ‘convenient’ sounds. The values of the levels correspond to a speed of 20 km/hour (the speed used for the listening test).

Figure 3 shows the spectrogram of two examples of QV sounds with a constant speed of 20 km/hour, with amplitude modulation, which results in alternating brighter and darker vertical bands. The levels of the factors for these examples are (A3, B4, C3, D4, E2, F4) (left) and (A2, B3, C4, D2, E3, F3) (right). For the first one, this corresponds to:

- (i)  $a_1 = 0.4$ ,  $a_2 = 0.8$ ,  $a_3 = 0.2$  and  $a_4 = 0.4$  (from A3 and C3);
- (ii)  $f_1 = 200$  Hz,  $f_2 = 300$  Hz,  $f_3 = 200$  Hz and  $f_4 = 600$  Hz (B4);
- (iii) Five harmonics for the motor and five for the chord (D4);

**Table 1.** Description of the design factors manipulated by the IGA. The values correspond to a speed vehicle of 20 km/hour.

Design factor	Level 1	Level 2	Level 3	Level 4
A: motor/chord proportion	$a_2 = 0$	$\frac{a_1}{a_2} = 2$	$\frac{a_1}{a_2} = \frac{1}{2}$	$a_1 = 0$
B: fundamental/centre frequency	$f_1 = 80$ Hz $f_2 = 120$ Hz $f_3 = 80$ Hz $f_4 = 240$ Hz	$f_1 = 120$ Hz $f_2 = 180$ Hz $f_3 = 120$ Hz $f_4 = 360$ Hz	$f_1 = 160$ Hz $f_2 = 240$ Hz $f_3 = 160$ Hz $f_4 = 480$ Hz	$f_1 = 200$ Hz $f_2 = 300$ Hz $f_3 = 200$ Hz $f_4 = 600$ Hz
C: harmonic components/noise proportion	$a_1 = a_2 = 0$	$\frac{a_3+a_4}{a_1+a_2} = 2$	$\frac{a_3+a_4}{a_1+a_2} = \frac{1}{2}$	$a_3 = a_4 = 0$
D: number of harmonics	Motor: 1 Chord: 1	Motor: 2 Chord: 2	Motor: 3 Chord: 3	Motor: 6 Chord: 6
E: amplitude modulation frequency	$f_m = 0.5$ Hz	$f_m = 2$ Hz	$f_m = 5$ Hz	$f_m = 10$ Hz
F: amplitude modulation ratio	$m = 0\%$	$m = 17\%$	$m = 33\%$	$m = 50\%$



**Figure 3.** Spectrogram of the complete sounds synthesised with parameters (A3, B4, C3, D4, E2, F4) (left) and (A2, B3, C4, D2, E3, F3) (right) with a constant speed vehicle (addition of the four synthesised components, with amplitude modulation).

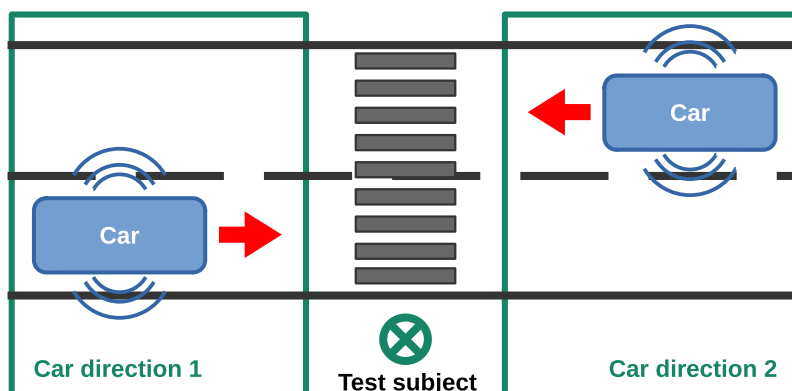
(iv)  $f_m = 2$  Hz and  $m = 50\%$  (E2 and F4).

For the second one, this corresponds to:

- (i)  $a_1 = 0.89, a_2 = 0.45, a_3 = 0$  and  $a_4 = 0$  (from A2 and C4);
- (ii)  $f_1 = 160$  Hz,  $f_2 = 240$  Hz,  $f_3 = 160$  Hz and  $f_4 = 480$  Hz (B3);
- (iii) One harmonic for the motor and five for the chord (D2);
- (iv)  $f_m = 5$  Hz and  $m = 33\%$  (E3 and F3).

### 3.2. Listening test scenario

The experiment aims at evaluating the unpleasantness and the detectability of QV sounds. To this end, the following scenario is considered: a pedestrian, standing on the sidewalk of a street, waits before crossing (see Figure 4). A QV may pass by,



**Figure 4.** Passing-by scenario for the listening test: pedestrian located on the sidewalk of a street.

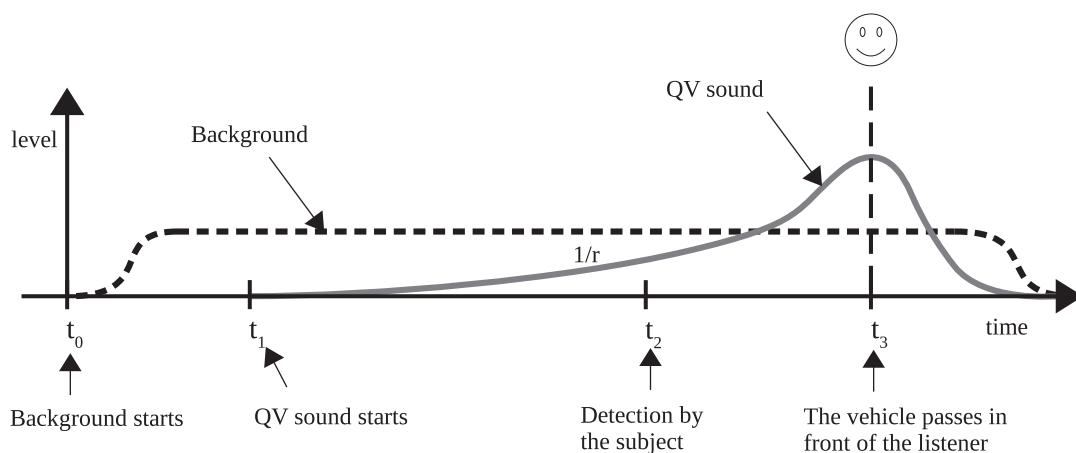
coming either from the right or from the left. The listener is static, and must indicate when he/she detects the QV. For all the passages, the speed of the car is kept constant (20 km/hour) and the direction of the car is randomly chosen as left or right.

To improve the realism of the test, the car sound is mixed with a urban environment background, made from a stereo street recording of a busy intersection in Paris, France. To be used as background noise, the soundscape must be amorphous (Maffiolo 1999) and not contain any perceivable emergent event (horns, car passing etc.; Kerber & Fastl 2008). For this, distracting sounds and close vehicle sounds that could be mistaken for the QV warning sound to evaluate are edited out of the recording. To avoid the potential fatigue of the participant due to the repetition of the exact same background noise during the test, the part of the audio file selected (around 10 seconds) is randomly chosen among the 42-second long recording.

To increase the level of immersion of the listener, and to obtain a more realistic passing-by scenario, the following properties have been implemented for the design of the sound stimuli:

- (i) The sound level of the QV is modulated according to the vehicle/listener distance. The model used, based on acoustic theory, considers the QV as a monopole and provides a sound level inversely proportional to the distance to the listener ( $\frac{1}{r}$ ; see Figure 5).
- (ii) The Doppler effect is simulated with a shifting in frequency due to the moving source.
- (iii) The left/right panning of the QV sound is controlled in such a way that the source goes progressively from one canal (left or right, depending of the direction of the QV) to the other (right or left) according to the position of the vehicle. The amplitude panning is controlled using a sine law (Pulkki 2001).

For the sake of simplicity, the simulation does not include other sound sources, such as tire noise.



**Figure 5.** Timeline of the mix of the background and the quiet vehicle (QV) sound, with their respective amplitude-level evolution (the  $x$ -axis represents indifferently the time or the distance of the QV, given that the speed of the vehicle is constant). Note the asymmetry of the amplitude relatively to the listening point.

In order to reduce the duration of each evaluation, the attenuation function of the QV sound is asymmetrical (reduction in  $\frac{1}{r}$  once the car passed in front of the listener), as in Misdariis *et al.* (2013), that is, the attenuation is faster than the increase of the sound level. This choice does not affect the detectability of the QV sound, which always occurs in the approach phase. We assume that this asymmetry does not have any effect on the assessments of unpleasantness.

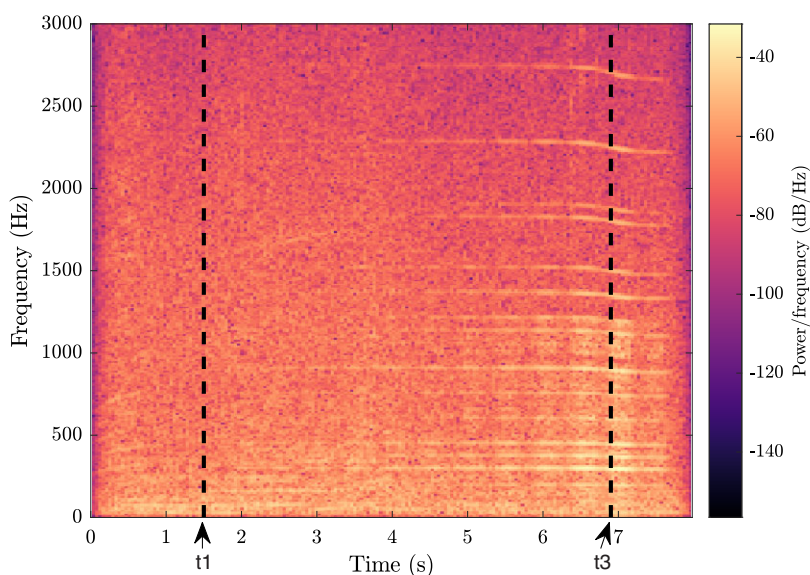
The scenario's timeline is shown in Figure 5. The total duration of each sound stimulus is around 10 seconds.

Figure 6 shows the spectrogram of the previous QV sound (A3, B4, C3, D4, E2, F4; Figure 3), now spatialised and mixed with the background noise. The masking of the warning sound by the background is illustrated by the fact that the horizontal stripes only appear some time after  $t_1$ . The shifting in frequency, due to the Doppler effect, is visible after time  $t_3$ , corresponding to the listener position. The sound can be listened to at <https://mathieulagrange.github.io/souaille2021interactive/>.

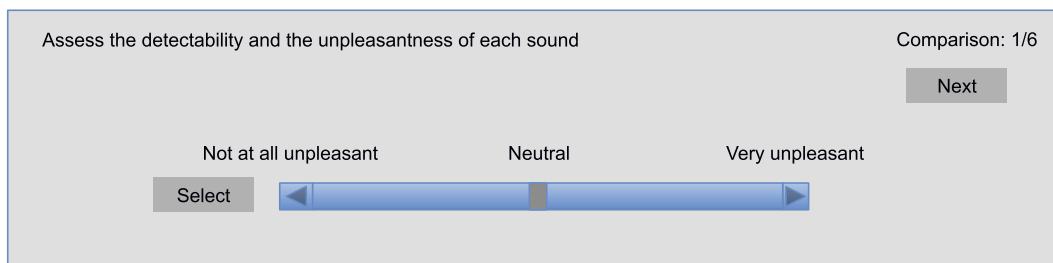
### 3.3. Test procedure and interface

The participants listened to the scenarios using computers and Beyerdynamics DT-990 headphones. The sound level was calibrated so that the background sound is around 69 dBA, when measured with a sound-level meter at the headphone's output. This level was chosen to be consistent with the dBA level measured during the recording of the background. The warning sound level relative to the background was manually adjusted to avoid having too many sounds detected too early or too late. The mean levels of the warning sounds ranged from 59 dBA to 80 dBA, with 95% of the sounds in the design space having values above 68 dBA (5th percentile).

The interface for the assessments of the QV sound is shown in Figure 7. After clicking on the 'Select' button, which launches the synthesis of the sound, participants had to strike the 'space bar' to start playing the sound. This corresponds to the definition of time  $t_0$  (see Figure 5). Next, they had to strike on the keyboard the



**Figure 6.** Spectrogram of an example of a quiet vehicle sound, spatialised and mixed with the background. Time  $t_1$  is the time at which the vehicle warning sound starts, and time  $t_3$  is the moment the vehicle passes in front of the listener. The horizontal stripes correspond to the harmonic content of the warning sound, progressively emerging from the background.



**Figure 7.** Interface for the assessment of the detectability and the unpleasantness of a quiet vehicle sound (structured rating scale).

‘a’ key as soon as they detect the QV coming from the left, or the ‘e’ key if it is coming from the right (French AZERTY keyboard). This strike allows the definition of the detection time  $t_2$ . To avoid habituation of the participant in the detection time (and detect inconsistent subjects), the starting time  $t_1$  of the QV sound in the mixture was variable, randomly chosen in the interval [1, 3] seconds. Of course, given this small interval, the event is highly predictable. A larger interval was not reasonable, as it would have increased the duration of the test. Hence, the protocol provides an estimate of the lower bound of the detection time, because it is clear that in real-life situation, when people do not wait for the arrival of a car, the detection time would be larger. This lower bound of the detection time is considered as representative of the detectability of the sound of the car.

The detection time is then given by

$$D_t = t_2 - t_1. \quad (7)$$

If the subject pressed a key before the vehicle warning sound is actually playing, that is, before  $t_1$ , the detection time of the sound is changed by default to an arbitrary value (average between the minimum and maximum possible values for the detection time  $[\frac{t_1+t_3}{2}]$ ), and a warning was recorded. If the car was not detected or detected too late, that is, after time  $t_3$ , the detection time was set to  $t_3$ , that is, the time at which the car passes in front of the subject, and a warning was recorded. If the subject made a mistake when assessing the arrival direction of the car, no action was taken, but a warning was recorded. This random change in the direction of the car is very important in the protocol to be able to detect ‘false alarms’ or ‘wrong detections’, in order to check the reliability of the assessments. The detection time can be converted into the distance to pedestrian using  $V$ , the speed of the car. The distance to pedestrian  $D_p$  at the time of the detection is then given by

$$D_p = (t_3 - t_2) \times V. \quad (8)$$

With the distance to pedestrian, a safety zone can be defined given the stopping distance of the car. The stopping distance of a car on a dry road is around 7.5 m at 20 km/hour. Thus, distances to pedestrian  $D_p$  lower than 7.5 m are considered as dangerous. This corresponds to a detection time  $D_t$  greater than 4.05 seconds.

After listening to the sound clip, the participants had to evaluate the unpleasantness of the sound on a continuous structured semantic scale going from ‘0’ (‘Not at all unpleasant’) to ‘10’ (‘Very unpleasant’) using a slider as shown in [Figure 7](#). To explain the semantic dimension of unpleasantness, the following information is given to the participants. ‘If the car passed by your house during a calm moment, how unpleasant would the sound be?’

Participants were able to replay the stimuli to assess the unpleasantness (as many times as required), but not to assess detectability. Indeed, they had already heard the sound and knew the direction of arrival of the car.

In the beginning of the test, the subjects were presented with a tutorial, so that they could familiarise themselves with the interface and understand which type of sounds they should pay attention to.

## 4. Experiment 1: IGA

### 4.1. Materials and method

For the first experiment, 32 students (16 males and 16 females) from the École Centrale de Nantes, France, with no reported auditory deficiencies, used the IGA procedure. They evaluated 11 generations of 9 sounds (99 sounds), which took approximately half an hour. Values of  $m_r = 0.7$ ,  $c_r = 0.25$  and  $s_r = 0.05$  were used for the IGA. A high mutation rate is chosen, to preserve diversity in spite of the small number of individuals per generation and to avoid premature convergence. At the end of the test, for each participant, the following information is available:

- (i) the set of Rank 1 sounds of the register at the last generation, that constitutes Pareto optimal solutions;
- (ii) all the sounds assessed during the 11 generations.



**Linear model of detectability and unpleasantness**

To study the influence of the design factors of the sounds on the detectability and the unpleasantness, a linear model without interactions is fitted to the data. For all participants, all sounds generated during the IGA experiments are used for the modelling (union of all the sounds assessed during the 11 generations). The model corresponds to a linear mixed model (similar to an analysis of variance [ANOVA]), with the six factors A, B, C, D, E and F with a fixed effect and the factor ‘subject’ with a random effect (Khuri, Mathew & Sinha 1998). The model is given by

$$y_{ijklmnop} = \mu + A_i + B_j + C_k + D_l + E_m + F_n + S_o + \varepsilon_{ijklmnop} \quad (9)$$

with:

- (i)  $y_{ijklmnop}$ : detection time, or unpleasantness rating, for the observation  $p$  of the sound  $(A_i, B_j, C_k, D_l, E_m, F_n)$  by participant  $S_o$ ;
- (ii)  $\mu$ : intercept;
- (iii)  $A_i$ : coefficient of the level  $i$  of factor  $A$ , with  $\sum_i A_i = 0$  (centred parameterisation). The other coefficients correspond obviously to factors B, C, D, E and F;
- (iv)  $S_o$ : coefficient of the subject  $o$ ;
- (v)  $\varepsilon_{ijklmnop}$ : error term,  $\varepsilon \sim N(0, \sigma^2)$ .

The determination coefficients  $R^2$  of the models are examined, together with the importance of the factors and their effect using the Fisher significance test. The percentage of importance  $I_j$  of factor  $A$  in the model is given by

$$I_A(\%) = \frac{\max_i(A_i) - \min_i(A_i)}{\sum_{M \in A,B,C,D,E,F} (\max_k(M_k) - \min_k(M_k))}. \quad (10)$$

Expressions are similar for the other factors.

Two models are fitted to the data, one for detection time and one for unpleasantness. An analysis of the parameters of the model is made in order to understand the main effects of the factors on the detectability and the unpleasantness.

**Definition of the ‘best’ individual sound  $IGA_{opt}^i$**

The following method has been set to define a unique QV sound, labelled  $IGA_{opt}^i$ , for each participant  $i$ . The first stage is to discard solutions of the Pareto set that are too extreme according to one of the objectives. Among the solutions of the individual Pareto set, solutions for which the detection time was above 4.05 seconds were rejected (detection time below the safety zone; see Section 3.3). Similarly, solutions with a very high unpleasantness relative to the unpleasantness range used by the participant were rejected. To do so, a min-max normalisation of the participant’s unpleasantness evaluations was performed, and the sounds in the upper third of the resulting range were withdrawn.

The second stage consists in the definition of a unique sound. With the remaining sounds, for each participant  $i$ , the TOPSIS method (Technique for Order of Preference by Similarity to Ideal Solution; Hwang & Yoon 1981) was used to select a unique optimal sound,  $IGA_{opt}^i$ . The first step of this method is to build a matrix  $x_{kj}$  from the remaining sounds’ objective values, where each row corresponds to a sound  $k$  and each column corresponds to an objective (in our case,  $j = 2$ ). Then, each column of  $x_{kj}$  is normalised by  $\sum_k x_{kj}^2$  and multiplied by a

weight. In our case, we give an equal weight of 0.5 to each objective. Two ideal solutions are then defined:

- (i) the positive ideal solution (PIS) which has, for each objective, the lowest value taken by the solutions in the considered group, meaning that  $PIS = \{\min(x_{k1}), \min(x_{k2})\}$ ;
- (ii) the negative ideal solution (NIS) which has, for each objective, the highest value taken by the solutions in the considered group, meaning that  $PIS = \{\max(x_{k1}), \max(x_{k2})\}$ .

After computing the Euclidean distance between each solution and the PIS and the NIS, respectively, called  $d_{PIS}$  and  $d_{NIS}$ , the following distance is computed:

$$distance = \frac{d_{NIS}}{d_{NIS} + d_{PIS}}. \tag{11}$$

The chosen optimal solution  $IGA_{opt}^i$  of participant  $i$  is the one that maximises this distance, which equals 1 if the sounds happen to be the PIS and 0 if it is the NIS.

This process allows the definition of individual optimal sounds  $IGA_{opt}^i$ , one for each participant  $i$ .

**Analysis of Optimalset, the set of Pareto optimal solutions**

The union, for all the participants, of all the individual Pareto solutions is formed. This set, labelled *Optimalset*, represents a selection of QV sounds that, from a perceptual point of view, make a satisfying trade-off between detectability and unpleasantness for the participants. To provide information that could be used as recommendations for a sound designer, an analysis of these sounds according to the most occurring factor-level combinations is conducted.

To draw design recommendations, the principle of the method is to consider the selection process of the designs made during the IGA experiment as a random process that depends on a discrete probability distribution. The set *Optimalset*, of size  $N$ , is a subset of the sample space  $\Omega$  (full factorial design). From the chosen designs in *Optimalset*, estimates of the parameters of the probability distribution can be calculated. And with these parameters, it becomes possible to make inferences and provide a probability score for any design of the design space.

*Joint probability.* Given the sample space  $\Omega$  (set of all possible designs of the design space), and the design variables  $X_i$ , ( $i = 1-6$ ) that describe the design, the first model that can be made is to assume that the choice of the designs in *Optimalset* depends on all the variables and all their possible interactions. In this case, the probability distribution of the selection process of any design  $d$  defined by the design variables  $X_i$ , ( $i = 1-6$ ),  $d = (X_1 = x_1, X_2 = x_2, \dots, X_6 = x_6)$  by the IGA experiments is given by the joint probability:

$$P(D = d) = P(X_1 = x_1, X_2 = x_2, \dots, X_6 = x_6) = \frac{card\{D \in Optimalset / D = d\}}{card(Optimalset)}, \tag{12}$$

where *card* represents the cardinality of a set (number of elements). For example, if a design is present once in *Optimalset*, its probability is  $P = \frac{1}{card(Optimalset)}$ .

If it is not chosen, its probability is  $P = 0$ . This six-dimension joint probability is not so interesting to make design recommendations because it is only able to

recommend designs that are present (and abundant) in *Optimalset*. To be able to make recommendations on the levels of the design variables  $X_i$ , it is necessary to make assumptions on the independence of the variables in the selection process.

*Marginal probability.* If we consider that the variables  $X_i$  ( $i = 1-6$ ) are mutually independent in the selection process (no interaction between them), then the probability distribution of the selection process of any design  $d = (X_1 = x_1, X_2 = x_2, \dots, X_6 = x_6)$  becomes

$$P(X_1 = x_1, X_2 = x_2, \dots, X_6 = x_6) = \prod_{i=1}^6 P(X_i = x_i). \quad (13)$$

When the variables are mutually independent, the joint probability is simply the product of the marginal probabilities, where

$$P(X_i = x_i) = \frac{\text{card}\{D \in \text{Optimalset} / X_i = x_i\}}{\text{card}(\text{Optimalset})}. \quad (14)$$

In this case, the design with the largest probability, that is, the one that should be recommended, is the design with the most occurring level for each variable. Of course, the mutual independence of all the variables is a very strong assumption that only holds if there is no interaction between the variables in the selection process (in the perception of participants). This is rather unlikely in design where the global assessment of a product may be different to the sum of the assessments of each of its variables (Sylcott, Michalek & Cagan 2015).

*Independence checking of the variables.* Between the two previous methods that have limited applicability for making design recommendations, it is interesting to propose a model that is based on assumptions concerning the independence of the variables in the selection process that can be checked. Our proposal is to check, with a statistical test, the independence of any pairs of variables in *Optimalset*. We propose to use the chi-square independence test to determine whether there is a significant association between two qualitative variables. For example, suppose that the pairwise independence test shows that the two groups of variables  $\{X_1, X_2, X_3\}$  and  $\{X_4, X_5, X_6\}$  are independent. Then, from this information, it is possible to simplify the expression of the joint probability and get the probability distribution of the selection process for any design  $d = (X_1 = x_1, X_2 = x_2, \dots, X_6 = x_6)$ , given by

$$P(X_1 = x_1, X_2 = x_2, \dots, X_6 = x_6) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \cdot P(X_4 = x_4, X_5 = x_5, X_6 = x_6), \quad (15)$$

where

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{\text{card}\{D \in \text{Optimalset} / X_1 = x_1, X_2 = x_2, X_3 = x_3\}}{\text{card}(\text{Optimalset})}, \quad (16)$$

$$P(X_4 = x_4, X_5 = x_5, X_6 = x_6) = \frac{\text{card}\{D \in \text{Optimalset} / X_4 = x_4, X_5 = x_5, X_6 = x_6\}}{\text{card}(\text{Optimalset})}. \quad (17)$$

It is then possible to calculate the probabilities of all the designs  $d = (X_1 = x_1, X_2 = x_2, \dots, X_6 = x_6)$  of the design space.

It is important to note that if all the variables are dependent (conjunct graph), the recommendations are simply the list of designs of *Optimalset*.

*Recommended designs.* A ranking of the design space by decreasing probability allows the definition of the ‘best’ designs, that is, the designs with the largest probabilities, to be recommended. Some of them are of course present in *Optimalset*, but it is likely that designs that are not present in *Optimalset* will get a high probability, and be interesting for the design problem. These designs may possess interesting characteristic combinations that explain their presence in the Pareto set. A set of eight sounds (labelled *reco1* to *reco8*) that get the highest probability score is proposed as recommended designs. They will be compared to other designs in Experiment 3.

#### **Outliers detection procedure**

Three indicators were considered to assess the performances of the participants in the detection task for the different experiments:

- (i) the direction error rate *DER*: percentage of stimuli detected with wrong direction;
- (ii) the early detection rate *EDR*: percentage of stimuli detected before  $t_1$ ;
- (iii) the late detection rate *LDR*: percentage of stimuli detected after  $t_3$ .

Limit values were defined for the different indicators, by considering the difficulty of the task and possible careless mistakes, inevitable given the relatively high cognitive load required for the experiments. The objective of these limits is not to select the participants, but to discard the dilettantes that did not provide the necessary commitment in the experiment. For these reasons, the limits are relatively large. So that the ratings of a participant are valid, it is necessary to have  $DER < 40\%$ ,  $EDR < 20\%$  and  $LDR < 85\%$ . Otherwise, their data were withdrawn from the study.

## **4.2. Results**

#### **Outlier detection**

Table 2 shows the average performance indicators (with the standard deviation between brackets) of the participants for the three experiments (Direction error rate  $\overline{DER}$ , early detection rate  $\overline{EDR}$  and late detection rate  $\overline{LDR}$ ). For Experiment 1, these average indicators show acceptable performances of the participants, indicating that the protocol is correctly designed and does not require outstanding abilities from the participants. The direction error rate (7.1%) remained weak and can be explained by haste mistakes of the participants. One participant made many wrong direction detections (above 40%), which is interpreted as a sign of an inability to use the interface correctly. This is probably due to some misunderstandings of the instructions. For those reasons, the data from this participant (several standard deviations away from the average) were not considered for the analysis. The early detection rate was very low (0.4%), which is a sign that the participants waited for the car and did not rush the test. The large average late detection rate (17.6%) indicates that some sounds are particularly hard to detect for

**Table 2.** Average participants performance rates and number of participants  $n$  not meeting the control limits, for the three experiments. Standard deviations (SD) of the rates are indicated between brackets.

	Direction error		Early detect		Late detect	
	$\overline{\text{DER}}$	$n \geq 40\%$	$\overline{\text{EDR}}$	$n \geq 20\%$	$\overline{\text{LDR}}$	$n \geq 85\%$
Experiment 1	7.1% (10.1)	1	0.4% (2.0)	0	17.2% (25.3)	3
Experiment 2	2.5% (3.2)	0	0% (0)	0	6.1% (9.1)	0
Experiment 3	3.8% (4.1)	0	0.2% (0.7)	0	10.0% (11.2)	0

some participants. Three participants had very high late detection rate (above 85% – due probably to a weak involvement in the experiment). For this reason, their data (several standard deviations away from the average) were excluded from the analysis. In summary, four participants among the 32 were withdrawn from the analysis due to low performance rates in their detection, leading to 28 valid participants.

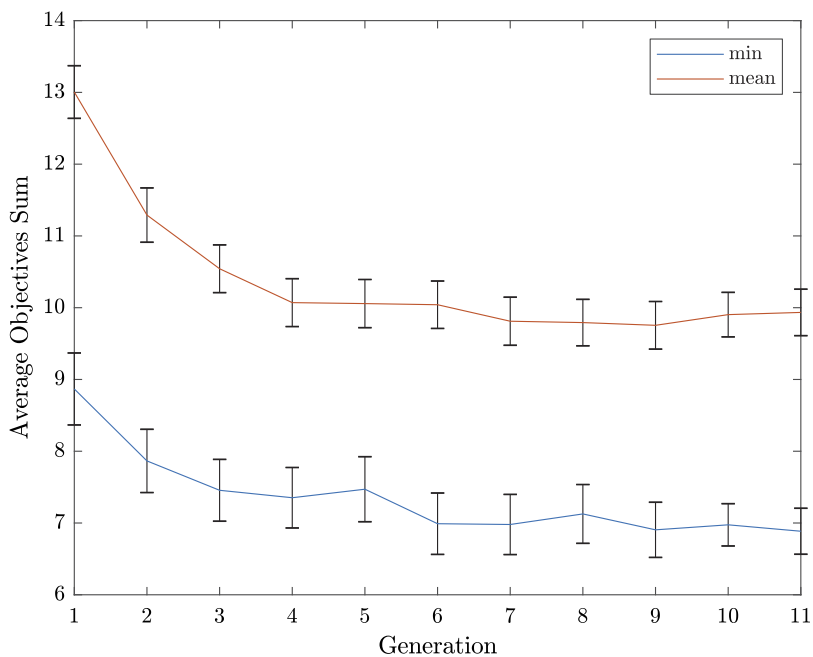
**Convergence of IGA**

The sum of the two objectives was examined in order to show the convergence of the solutions across the different generations. Before summing, the value of the detection time was scaled so that its range matched the one of the unpleasantness. Please note that the IGA do not directly operate on this sum as it is a multiobjective optimisation. That being said, we believe that this reduction to a single objective is a convenient way to monitor the behaviour of the IGA.

Figure 8 shows the sum averaged over all participants plotted for two conditions: the mean value for the solutions of a generation and the minimum value for a generation. On average, the objectives’ sum is decreasing over the generations, which shows an improvement of the proposed sounds with regard to the design problem. This is a sign of the reliability of the experimental protocol for the assessment of the detection time and the unpleasantness, and a correct tuning of the IGA parameters.

**Analysis of the effect of the factors**

The results of the two linear models between the sound characteristics (detectability and unpleasantness) and the six design factors are given in Table 3. All the assessments made by the valid participants ( $28 \times 99 = 2772$  observations) during the IGA experiments are used for the modelling. For the two models, the determination coefficient  $R^2$  is not large (43% and 46%), indicating that the models are approximate and gather only a general trend in the data. They thus cannot be used to make accurate predictions of the responses (recommendations). Nevertheless, all the factors are significant in the model ( $p < 0.001$ ), and they are therefore useful to explain the variability in the two responses. The importance of the factors is rather similar (around  $100/6 = 16.6\%$ ), except Factor E (amplitude modulation frequency), which has a weaker influence on detectability (7.1%) and unpleasantness (6.4%), and Factor B (fundamental/centre frequency), which has a larger influence (34.3% and 27.2%).



**Figure 8.** Average value of the sum of the two objectives versus generations, with the standard error. The value of the detection time has been scaled so that its range matches the one of the unpleasantness.

The values of the coefficients of the models (graph of the main effects) are presented in Figure 9 for unpleasantness and detectability. From these graphs, it is clear that there is an opposition between detectability and unpleasantness: the general trend is that to have a ‘pleasant’ sound, the levels of all the factors must be low (Level 1), and to have a detectable sound, the levels of all the factors must be high (Level 4). The lower the level’s modality, the more pleasant the sounds, and the higher the level’s modality, the more detectable the sounds. Factor E has a little more complex effect for Level 4 (Level 4 is not the most unpleasant and the most detectable). This analysis confirms the trade-off between detectability and unpleasantness that is mentioned in other studies (Parizet *et al.* 2013; Lee *et al.* 2017). Furthermore, an examination of the values of the different levels shows that the results are in accordance with previous studies: to be ‘not unpleasant’, the sound must resemble an ICE sound (A1), be low pitched (B1 – low frequency), be made of noise (C1 – no harmonic sound; see the whine index in Lee *et al.* 2017) and with no modulation (F1); to be detectable, the sound must be high pitched (B4 – high frequency), with an harmonic content (A4, C4) and with modulation (F4).

However, this opposition between detectability and unpleasantness is only a general trend, revealed by a linear model that account for less than 50% of variance. Particular QV sounds with adapted combinations of factor levels may optimise at the same time detectability and unpleasantness, but this approximate model does not provide us with enough information to find these trade-off sounds. This motivates the need for an alternative method for making recommendations.

**Table 3.** Coefficients,  $p$ -value of the significance Fisher test and importance of the factor for the two linear models (unpleasantness and detectability)

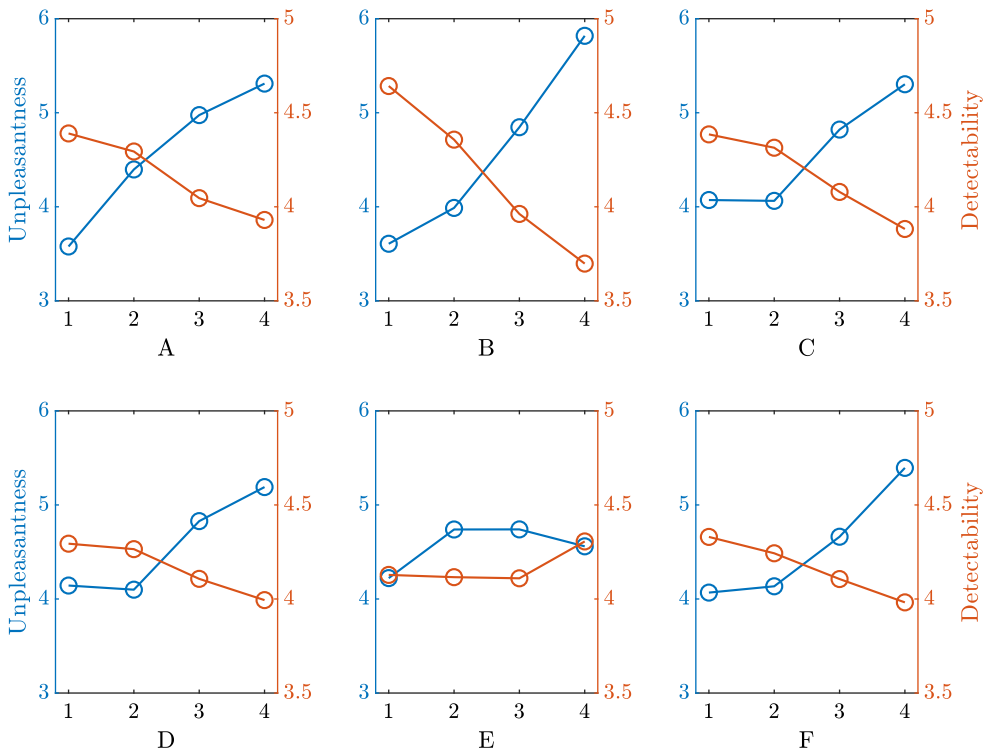
Factor	Level	Unpleasantness ( $R^2 = 43\%$ )			Detectability ( $R^2 = 46\%$ )		
		Coeff.	Signif.	I(%)	Coeff.	Signif.	I(%)
A	A1	-0.99			0.23		
	A2	-0.17			0.13		
	A3	0.41	$p < 0.001$	21.3%	-0.12	$p < 0.001$	16.7%
	A4	0.74			-0.24		
B	B1	-0.96			0.48		
	B2	-0.58			0.19		
	B3	0.28	$p < 0.001$	27.2%	-0.20	$p < 0.001$	34.3%
	B4	1.25			-0.47		
C	C1	-0.49			0.22		
	C2	-0.50			0.15		
	C3	0.26	$p < 0.001$	15.3%	-0.09	$p < 0.001$	18.3%
	C4	0.74			-0.28		
D	D1	-0.42			0.13		
	D2	-0.47			0.10		
	D3	0.26	$p < 0.001$	13.4%	-0.06	$p < 0.001$	10.9%
	D4	0.62			-0.17		
E	E1	-0.34			-0.04		
	E2	0.17			-0.05		
	E3	0.17	$p < 0.001$	6.4%	-0.06	$p < 0.001$	7.1%
	E4	-0.01			0.14		
F	F1	-0.50			0.17		
	F2	-0.43			0.08		
	F3	0.10	$p < 0.001$	16.3%	-0.06	$p < 0.001$	12.6%
	F4	0.83			-0.18		

**Analysis of the sounds of *Optimalset***

*One variable at once.* The *Optimalset* (union of the Pareto front of all the participants) counts  $N = 113$  sounds. All of them are Pareto optimal, representing different compromises between unpleasantness and detectability. For each participant, the Pareto front of the best-solution register contained 1–8 solutions. All designs are present once, except d1(A1 B4 C4 D4 E2 F1) present twice and d2(A2 B4 C4 D4 E1 F1) present four times. The occurrences of the levels of each factor in *Optimalset* are given in Table 4.

For example, for the Harmonic/noise proportion (C), the level C4 (no noise) is chosen 50 times. To define the variables subjected to the most consensual choice concerning their levels, a multinomial goodness-of-fit test of the distribution of the occurrences is carried out. Results are presented in Table 4. Only two factors





**Figure 9.** Graph of the coefficients of the linear mixed models of unpleasantness and detectability for the six factors A, B, C, D, E and F.

**Table 4.** Occurrences of the levels of each factor in *Optimalset*

	A	B	C	D	E	F
Level 1	35	14	19	27	39	36
Level 2	20	36	21	28	23	33
Level 3	31	33	23	25	28	24
Level 4	27	30	50	33	23	20
Multinomial test signif.	N.S.	Sign.*	Sign.**	N.S.	N.S.	N.S.

Abbreviation: N.S.: not significant.

\* $p < 0.05$

\*\* $p < 0.01$

(B [frequency] and C [Harmonic/noise proportion]) obtain occurrences significantly different from a random distribution at the 5% level. For the frequency of the sounds, the level B1 (low frequency) is under-represented (size = 14). For the Harmonic/noise proportion, the level C4 (broad band noise absent) is over-represented (size = 50). For the whole group of participants, according to a majority compromise, it seems necessary to avoid low frequencies and broadband noise for the high detectability and low unpleasantness of QV sounds. For the other variables, it is not possible to make recommendations with this simple sorting one

**Table 5.** Pairwise comparison matrix of the chi-square test of independence (*p*-value)

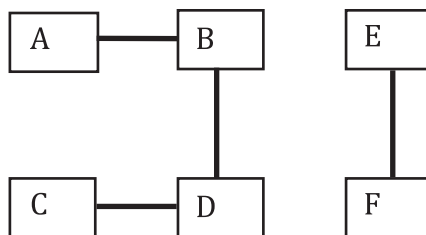
	A	B	C	D	E	F
A	/	<b>0.002</b>	0.003	0.953	0.264	0.393
B		/	0.015	<b>0.001</b>	0.052	0.109
C			/	<b>0.001</b>	0.53	0.466
D				/	0.219	0.026
E					/	<b>0.001</b>
F						/

variable at once, either because there are groups of subjects with different views on these variables, or because there is no clear influence of each variable alone on the compromise (interactions). These results are in line with the models proposed in the previous section. Given that the two objectives are conflicting for all the variables, there is no obvious optimum when one considers each variable alone.

*Independence test of the factors in Optimalset.* With the definition of the *N* designs of *Optimalset*, contingency tables of all pairs of variables are formed. The results of the chi-square test of independence (*p*-value) are given in Table 5. The *p*-values corresponding to a rejection of the independence are presented in bold (a Bonferroni correction is applied to deal with the multiple comparisons problem – threshold value of  $0.05/15 = 0.003$ , where 15 is the number of pairs). For the nonsignificant pairs, the test shows that the dependence in the sample is too weak to distinguish it from independence. From this table, the corresponding dependence graph can be drawn (Figure 10).

The graph shows that two groups of mutually independent variables can be considered in *Optimalset*: {*A, B, C, D*} and {*E, F*}. These indicate that there might be some interaction effect between Factors A, B, C and D on the performance of a sound, when considering both objectives. Likewise for Factors E and F. In other words, it signifies that it is not possible to make recommendations for the level of a single variable in the sets {*A, B, C, D*} or {*E, F*}. All the variables in the set must be considered. The empirical probability law of the selection process is then given by

$$P(A = A_i, B = B_j, C = C_k, D = D_l, E = E_m, F = F_n) \approx P(A = A_i, B = B_j, C = C_k, D = D_l) \cdot P(E = E_m, F = F_n). \tag{18}$$



**Figure 10.** Graph of the dependency between the factors.

**Table 6.** Definition of the eight recommended designs (design variables' levels). The dashed line highlights the independence between  $\{A, B, C, D\}$  and  $\{E, F\}$ .

	A	B	C	D	E	F	Probability score	Presence in Optimalset
<i>reco1</i>	A3	B2	C2	D2	E1	F1	0.58%	No
<i>reco2</i>	A1	B4	C4	D4	E1	F1	0.58%	Yes
<i>reco3</i>	A2	B4	C4	D4	E1	F1	0.58%	Yes
<i>reco4</i>	A3	B2	C2	D2	E3	F2	0.5%	Yes
<i>reco5</i>	A1	B4	C4	D4	E3	F2	0.5%	Yes
<i>reco6</i>	A2	B4	C4	D4	E3	F2	0.5%	No
<i>reco7</i>	A1	B2	C4	D2	E1	F1	0.46%	No
<i>reco8</i>	A4	B3	C4	D2	E1	F1	0.46%	No

From this expression, the probability scores of all the design samples of the full factorial can be calculated, based on the probability score of each group of variables.  $P(A = A_i, B = B_j, C = C_k, D = D_l)$  is computed as the number of designs in *Optimalset* for which  $A = A_i, B = B_j, C = C_k$  and  $D = D_l$ , divided by the number of designs in *Optimalset*. Likewise for  $P(E = E_m, F = F_n)$ . Designs with the largest probability (see Table 6) can be recommended as their factor-level combinations constitute attractive characteristics for the Pareto set. They are considered as relevant design solutions to the problem. Because the score contributions from the variable groups  $\{A, B, C, D\}$  and  $\{E, F\}$  are computed independently, some designs might have a high score even though they are not in *Optimalset*.

For the remainder of the paper, the independence between  $\{A, B, C, D\}$  and  $\{E, F\}$  will be indicated by a vertical separator in design descriptions (e.g., [A1 B4 C4 D4 | E2 F1] for *design1*).

*Design recommendations.* The eight designs with the largest probability scores (*reco1* to *reco8* – ranked by decreasing probability) are presented in Table 6. Half of them are presented in *Optimalset*. They present combinations of variables that may be interesting for the design problem. It is interesting to mention that different levels of the factors are present in the recommended designs: no single level presents a paramount advantage for the design problem. It is instead the combination of different levels that constitute an interesting sound. This is in agreement with the results of the models of unpleasantness and detectability, which showed that no recommendation can be made independently on each factor because of the conflicting objectives.

## 5. Experiment 2: individual validation

### 5.1. Materials and method

One week after Experiment 1 (time to analyse the results), the second experiment was proposed to the participants of Experiment 1. Twenty-four out of the 32 previous subjects re-evaluated their own optimal solution  $IGA_{opt}^i$ , as well as two

*designed* sounds (*design1* and *design2*) and four random sounds (*rand1*, *rand2*, *rand3* and *rand4*).

The *designed* sounds were proposed by a skilled person having a thorough knowledge of the possibilities offered by the synthesiser (T.S., one of the co-authors). These designs were intended to be good compromises between unpleasantness and detection time while having a different timbre with respect to one another. They were designed by tuning the parameters of the synthesiser while listening to its output, without prior information on the preferences of the test group. This allowed the designer to develop an intuition of how the design parameters related to his own perception of unpleasantness and detectability. The designer tried to find areas of the design space where a small change in parameters would make the sound either too unpleasant or not detectable enough. The sound *design1* is high pitched, with a small number of harmonics, no noise content and no amplitude modulation. Its parameter values are (A2 B4 C4 D2 | E1 F1). The sound *design2* is pitched slightly lower, with just noise and fast, low amplitude modulation. Its parameter values are (A1 B3 C1 D1 | E4 F2).

In addition to these two designed sounds, four *random* sounds (*rand1*, *rand2*, *rand3* and *rand4*) are generated by a random generation in the design space based on a Latin Hypercube Sampling. Those sounds are taken as baseline to judge the relative performances of the other sounds, *designed* and *IGA*. The sounds can be listened to at <https://mathieulagrange.github.io/souaille2021interactive/>.

All subjects evaluated the same *designed* and *random* sounds, together with their own optimum  $IGA_{opt}^i$ . All the assessments were repeated four times to stabilise variance. Thus, each participant had to evaluate  $7 \times 4 = 28$  sounds, presented one at a time (one per page in the interface), in a random order. For each evaluation, the interface was identical and there was no indication regarding which sound is being evaluated.

First, in order to assess the ability of subjects to perceive differences (intra-individual variability compared to inter-sounds variability), a one-way ANOVA (with the factor type of sounds) was done for each subject and each objective (omnibus test).

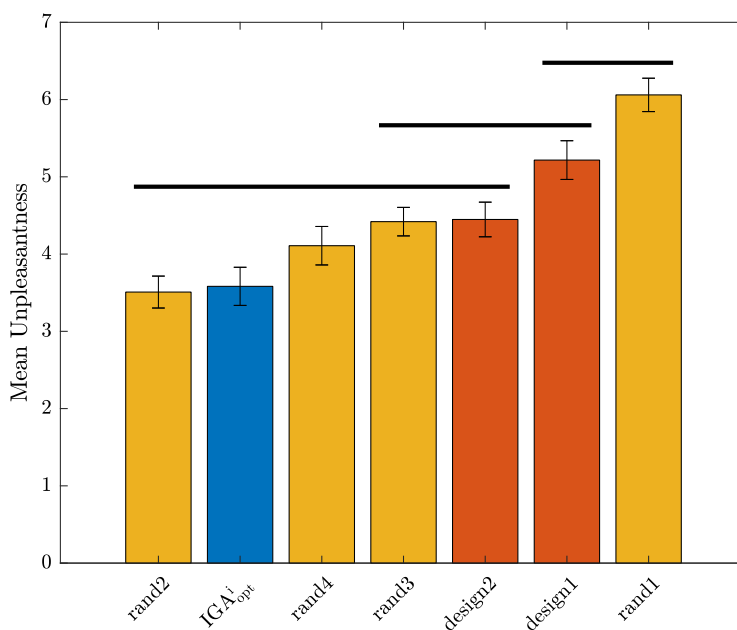
The sounds were next compared using two criteria:

- (i) Their membership to the Pareto front, based on their mean unpleasantness and detection time. This does not include any statistical test.
- (ii) The statistical significance of the pairwise differences between the mean unpleasantness values, for all pairs of sounds. Likewise for the detection time. Tukey's HSD (Honestly Significant Difference) test was used to assess these differences for each pair of sounds among  $IGA_{opt}^i$ , *design1*, *design2*, *rand1*, *rand2*, *rand3* and *rand4*.

## 5.2. Results

As illustrated in [Table 2](#), there was no outlier for this experiment. Therefore, all the participants were considered as valid.

The ANOVA performed on each subject's unpleasantness and detectability ratings showed a significant effect ( $p < 0.05$ ) of the sound on unpleasantness, for all but two subjects. Almost all the subjects were able to overall discriminate the sounds. For detectability, a significant effect was found for only 8 out of the



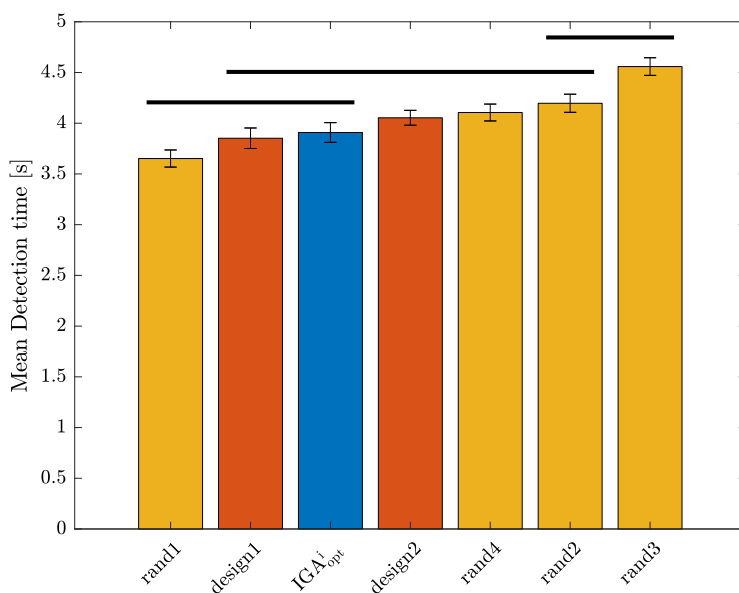
**Figure 11.** Experiment 2: average unpleasantness and results of the post hoc analysis for the different quiet vehicle sounds. The black horizontal lines show groups means that are not significantly different (Tukey's HSD test [ $p > 0.05$ ]).

24 subjects. The discrimination was then weaker for detectability. We believe that this is because the detectability differences between the sounds of Experiment 2 were too close, for certain subjects, to their just noticeable difference.

The mean evaluations of detection time and unpleasantness for all the participants and all the repetitions are computed for the following sounds:  $IGA^i_{opt}$ ,  $design1$ ,  $design2$ ,  $rand2$ ,  $rand2$ ,  $rand3$  and  $rand4$ . The average scores and their standard error are presented in Figure 11 for unpleasantness and Figure 12 for detection time. Results of Tukey's HSD multiple comparison tests are also presented in these figures (significant threshold:  $p = 0.05$ ): results are presented with bold lines connecting the sounds. When sounds are connected, pairs are not significantly different, whereas they are when not connected.

According to unpleasantness (Figure 11), the sound  $IGA^i_{opt}$  is the second least unpleasant. There is no significant difference within three groups of sounds: ( $rand2$ ,  $IGA^i_{opt}$ ,  $rand4$ ,  $rand3$ ,  $design2$ ), ( $rand3$ ,  $design2$ ,  $design1$ ) and ( $design1$ ,  $rand1$ ). Even if  $IGA^i_{opt}$  is not significantly less unpleasant than all the other proposals (particularly the sound  $rand2$  that is judged as the least unpleasant), it is among the least unpleasant sounds. In particular, it is significantly less unpleasant than the sound  $design1$ .

According to detection time (Figure 12), the sound  $IGA^i_{opt}$  is the third most detectable sound. There is no significant difference within three groups of sounds: ( $rand1$ ,  $design1$ ,  $IGA^i_{opt}$ ), ( $design1$ ,  $IGA^i_{opt}$ ,  $design2$ ,  $rand4$ ,  $rand2$ ) and ( $rand2$ ,  $rand3$ ). Again,  $IGA^i_{opt}$  is not significantly more detectable than all the other proposals, but it belongs to the three most detectable sounds. In particular, it is as detectable as the two designed sounds  $design1$  and  $design2$  (no significant difference).

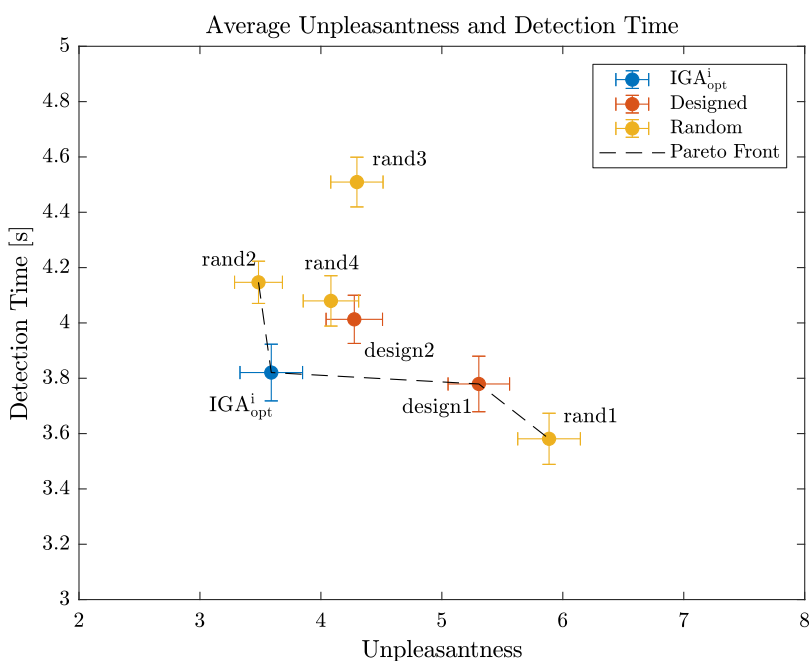


**Figure 12.** Experiment 2: average detection time and results of the post hoc analysis for the different quiet vehicle sounds. The black horizontal lines show groups means that are not significantly different (Tukey's HSD test [ $p > 0.05$ ]).

If we consider a stopping distance of 7.5 m at 20 km/hour, the detection time has to be below 4.05 seconds in order to avoid collision. This means that the sounds *rand4*, *rand2* and *rand3* do not allow the QV to stay in a safety zone with regard to detectability for blind people. The sound  $IGA^i_{opt}$  is in the safety zone, whereas the sound *design2* is close to the limit.

To examine the efficiency of the sounds according to the two objectives together, Figure 13 presents the positions of the sounds in the plane of the two objectives, with the corresponding standard errors. The results show that the individual sound of each participant  $IGA^i_{opt}$  makes on average an excellent compromise between the two objectives. Even if the sound  $IGA^i_{opt}$  does not Pareto dominate all the other sounds, it belongs to the Pareto front, with the sounds *rand2*, *design1* and *rand1*. The sound *rand2* is not unpleasant, but the price to pay is a large detection time (beyond the safety zone). Conversely, *rand1* is very detectable, but the most unpleasant.  $IGA^i_{opt}$  Pareto dominates the sound *design2*, whereas it can be considered as equivalent (the same rank) to the sound *design1*. In conclusion,  $IGA^i_{opt}$  constitutes an interesting trade-off between the two objectives, as well as the sound *design1*. The *random* sounds are either Pareto dominated by  $IGA^i_{opt}$ , or too extreme on one of the objectives to constitute relevant proposals.

This result validates the efficiency of the IGA experiment for the design of sounds, at least at the individual level. We can then give a positive response to question Q1: Sounds designed with the IGA experiment and selected with the TOPSIS method are interesting candidates for the design problem, comparable in terms of efficiency to sounds designed by a designer.



**Figure 13.** Experiment 2: scatterplot of the average performances of the different quiet vehicle sound categories (*IGA*, *Designed*, *Random*) according to the two objectives: unpleasantness and detection time. The dashed line indicates the Pareto front.

## 6. Experiment 3: validation of the recommendation method

### 6.1. Materials and method

In this experiment, a second panel of subjects (different from the previous panel) is asked to assess the detectability and unpleasantness of different QV sound proposals, including the recommended designs obtained with the method described previously (see the ‘Recommended designs’ and ‘Design recommendations’ sections). Seventeen students (14 males and 3 females) from the École Centrale de Nantes, France, with no reported auditory deficiencies, participated to the test with the protocol described previously. Eighteen QV sounds were proposed to the evaluation:

- (i) Eight sounds (*rand1* to *rand8* – category *Random*) randomly defined in the experimental space. *rand1* to *rand4* are the same random sound as in Experiment 2.
- (ii) Two sounds (*design1* and *design2* – category *Designed*), designed by a sound designer with instructions for ‘good detectability’ and ‘low unpleasantness’. These sounds are the same as those of Experiment 2.
- (iii) Eight sounds (*reco1* to *reco8* – category *Recommended*), recommended by the method described in the previous section (sounds with the largest probability according to the selection process).



To be able to assess the experimental variability in the assessments, four repetitions of each sound were proposed. In total, each participant had to assess 72 ( $18 \times 4$ ) QV sounds, proposed in a randomised order. The performances of each pair of sounds were compared using Tukey's HSD Multiple Comparison test. Just like for Experiment 2, the intra-individual variability was assessed with two ANOVAs for each subject, one for unpleasantness and one for detection time.

The sounds were compared using the same two criteria as for Experiment 2 (Membership to the Pareto front and multiple comparison tests).

## 6.2. Results

### **Outlier detection**

As indicated in [Table 2](#), there was no outlier for this experiment. Therefore, all the participants were considered as valid.

### **Intra-subject variability**

The one-way ANOVA performed on the subject's unpleasantness and detectability ratings showed a significant effect of the sound ( $p < 0.05$ ) for all but one subject in the case of the unpleasantness and all but another subject in the case of detectability. Almost all the subjects were then able to perceive significant differences between the sounds (omnibus test).

### **Multiple comparisons**

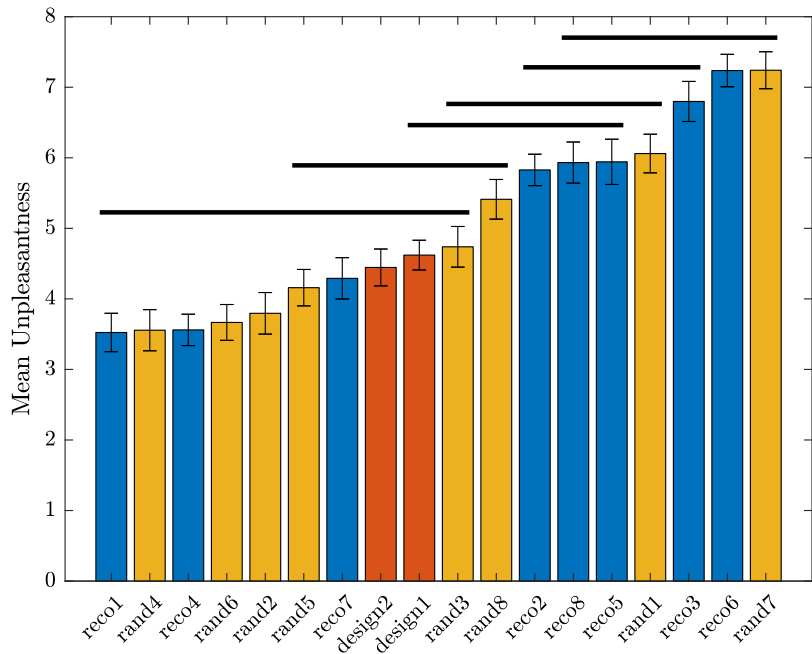
From the assessments of the participants according to detectability and unpleasantness, the average scores of detectability and unpleasantness of the 18 QV sounds were calculated. The average scores and their standard errors are presented in [Figure 14](#) for unpleasantness and [Figure 15](#) for detection time together with the Tukey's HSD multiple comparison test for every pair of sounds (significant threshold:  $p = 0.05$ ).

According to unpleasantness ([Figure 14](#)), three recommended designs (*reco1*, *reco4*, *reco7*) are present in the first group, which also includes the two designed sounds. The designed sounds are rather 'in the middle' in terms of unpleasantness. The random and recommended sounds are spread along the unpleasantness range.

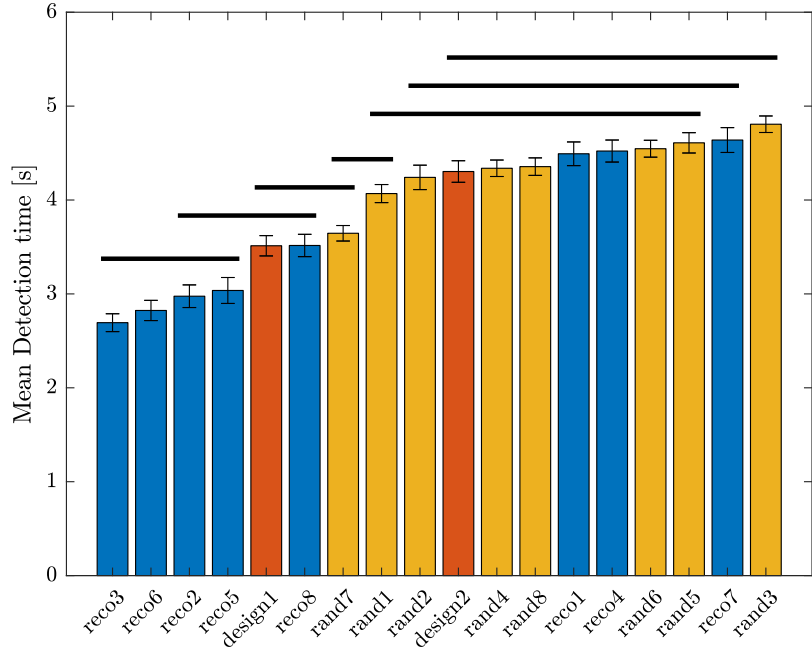
According to detection time ([Figure 15](#)), four recommended designs (*reco3*, *reco6*, *reco2*, *reco5*) are present in the first group. Again, the designed sounds are rather 'in the middle' regarding detectability. On the other hand, the random sounds are rather on the right side of the scale (large detection time), whereas the recommended sounds are on the left (low detection time). It is clear in [Figure 15](#) that the recommended designs are on average more detectable than the other sounds.

### **Average ratings**

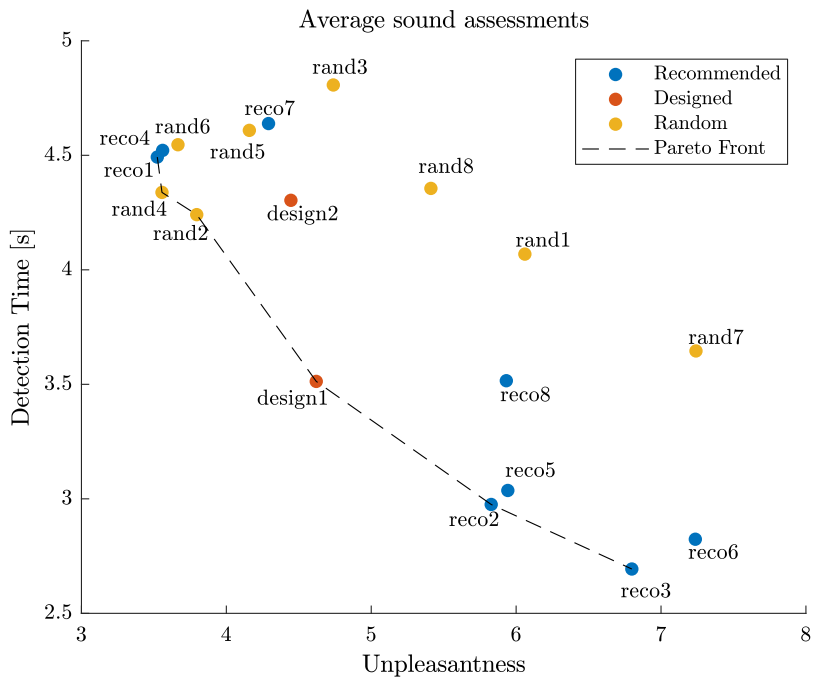
To have a more accurate view of the performances of the sounds according to the two objectives, and visualise the trade-off, [Figure 16](#) presents the average performances of each QV sound of the three categories (*Recommended*, *Designed*, *Random*) according to detection time and unpleasantness. All the sounds can be listened to at the following address, to give a better idea of their design: <https://mathieulagrange.github.io/souaille2021interactive/>.



**Figure 14.** Experiment 3: bar graph of the average value of the unpleasantness for the different quiet vehicle sounds. Nonsignificant differences between pairs of sounds ( $p > 0.05$ ) are linked with a horizontal line (Tukey's HSD multiple comparisons test).



**Figure 15.** Experiment 3: bar graph of the average value of the detection time for the different quiet vehicle sounds. Nonsignificant differences between pairs of sounds ( $p > 0.05$ ) are linked with a horizontal line (Tukey's HSD multiple comparisons test).

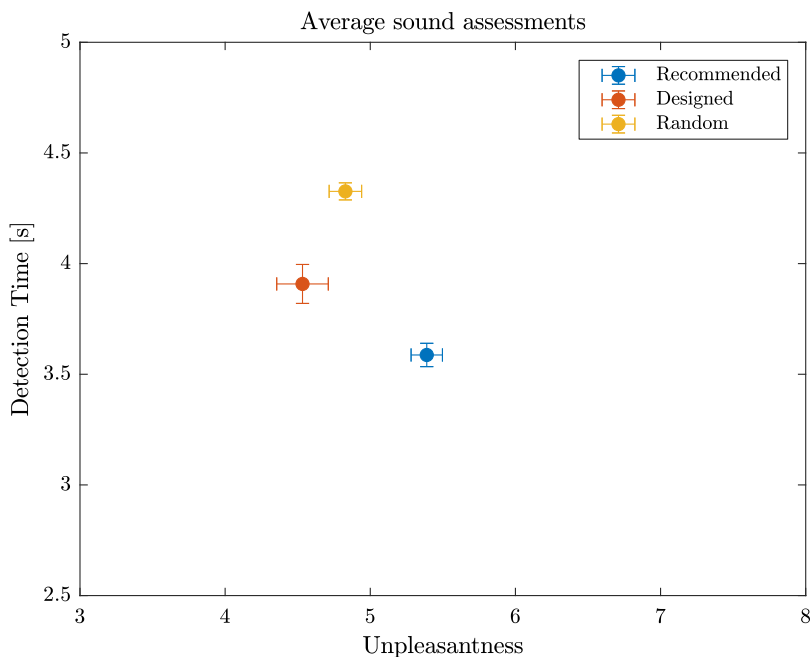


**Figure 16.** Experiment 3: scatterplot of the average performances of the different quiet vehicle sounds from each category (*Recommended*, *Designed*, *Random*) according to the two objectives, and visualisation of the Pareto front.

First, we can see that the relative position of the two designed sounds *design1* and *design2* is similar for Experiment 2 (Figure 13) and Experiment 3 (Figure 16): the sound *design1* is more unpleasant, but more detectable than *design2*. It belongs to the Pareto front in both cases. Similarly, the relative position of the sounds *rand1* to *rand4* is similar for both experiments. This result is a sign of the agreement of the two panels of participants and of the reliability of the experimental protocol.

Second, the Pareto front is made of three recommended sounds (*reco1*, *reco2*, *reco3*), one designed sound (*Design1*) and two random sounds (*rand2* and *rand4*). The three first recommended sounds (*reco1*, *reco2*, *reco3* – sounds with the largest probability) belong to the Pareto front. Furthermore, the three other recommended sounds (*reco4*, *reco5*, *reco6*) are very close to the Pareto front. The recommended sounds are rather well represented on the Pareto front.

Third, it is noticeable that the range of the random sounds according to unpleasantness is large, but very narrow for detectability, furthermore located near large detection times: it seems then unlikely to obtain detectable QV sounds by choosing them randomly. Conversely, the ranges of the *Recommended* sounds according to unpleasantness and detectability are large: the recommended sounds cover a large area of the Pareto front. The recommendations are various, which is an advantage for the definition of different sound prototypes. Furthermore, the most detectable QV sounds are recommended sounds (*reco2*, *reco5*, *reco3*, *reco6*). Sounds that are highly detectable seem to be very specific, given that neither random nor designed sounds obtain comparable performances.



**Figure 17.** Experiment 3: scatterplot of the average performances of the different quiet vehicle sound categories (*Recommended*, *Designed*, *Random*) according to the two objectives.

The two designed sounds obtain average performances, the sounds *design1* (Pareto efficient) being an interesting trade-off between the two objectives.

In conclusion, even if all the recommended sounds are not Pareto efficient, they present a large diversity of proposals and are close to the Pareto front (except *reco7* and *reco8*).

This result validates the efficiency of the recommendation method: recommended sounds are on average better than random sounds, and comparable to designed sounds, while offering different trade-offs.

This is illustrated by [Figure 17](#), showing the mean unpleasantness and detection time by sound category.

We can then give a rather positive response to question Q2: The recommendation method and the experimental protocol using the IGA allow the determination of very detectable QV sounds, as well as sounds with low unpleasantness.

## 7. Experiments summary

### 7.1. Method

The different stages for the definition and the analysis of the three experiments are summarised in [Figure 18](#). The flowchart presents the linking of the different stages, and the analysis of the results to address the research questions Q1 and Q2:

- (i) Experiment 2 allows the addressing of question Q1, with a comparison at the individual level of the performances of the IGA sounds for detectability and unpleasantness;

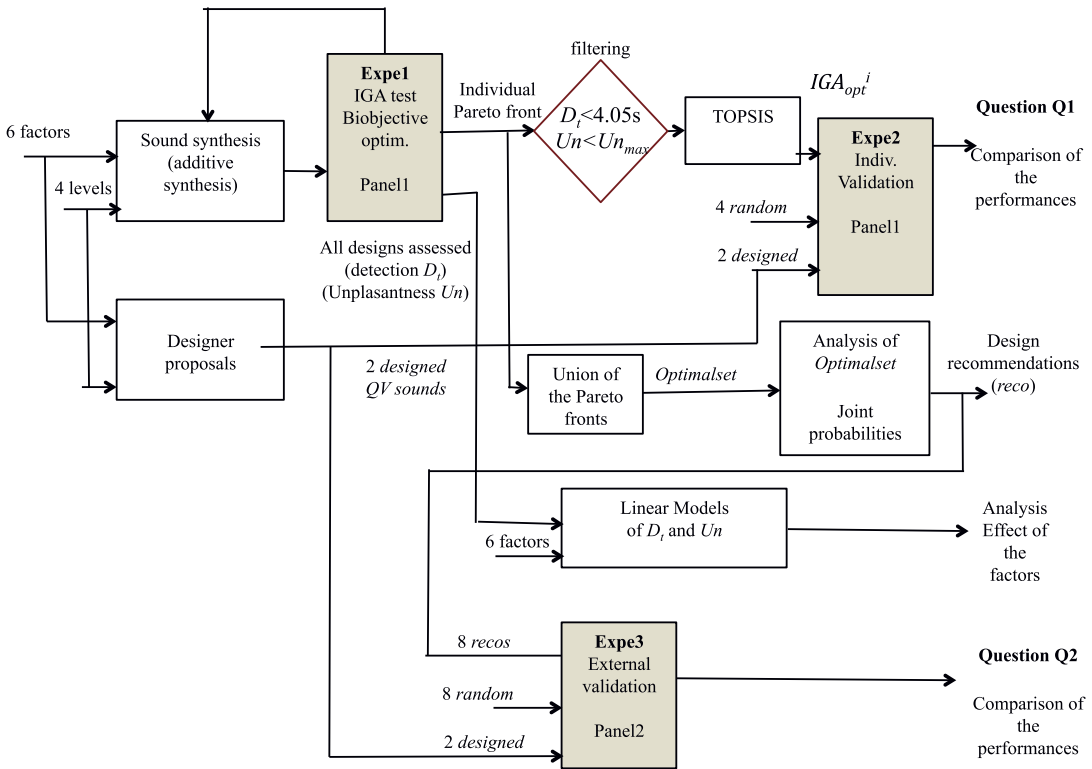


Figure 18. Overview of the flowchart for the analysis of Experiments 1–3.

- (ii) Experiment 3 allows the addressing of question Q2, with a comparison of the performances of the recommended sounds for detectability and unpleasantness with an external panel of participants.

## 7.2. Results

Figure 19 shows a simplified version of the flowchart of Figure 18, along with an overview of the keys results of Experiments 1–3.

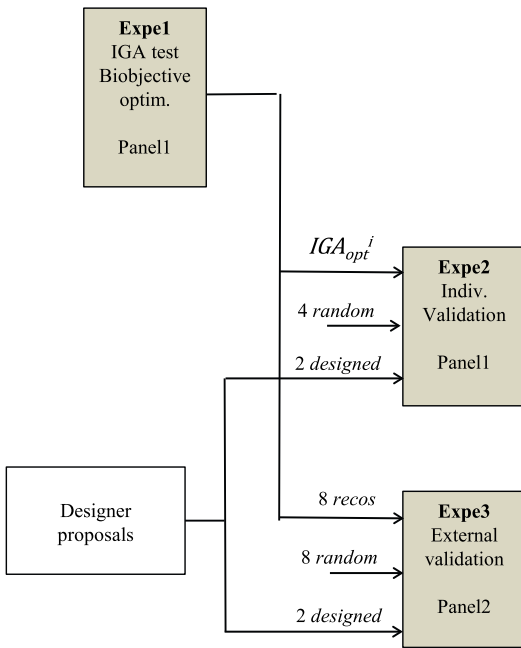
## 8. Analysis/discussion

This section is dedicated to the implications, in terms of design, of the outcomes of this study.

### 8.1. Sound level

Even though a study of the acoustical properties of the sounds is out of the scope of this paper, it is important to note that the sounds were not equalised in loudness during the experiments. Since perceived loudness is expected to be an important factor for unpleasantness and detectability, an analysis of the A-weighted equivalent continuous sound level  $L_{Aeq}$  of the evaluated warning sounds was done for

Results



- The evolution of the average sum of unpleasantness and detection time showed an improvement of the sounds' performances during the IGA process.
- The two linear mixed models showed an opposition between unpleasantness and detection time.
- A unique best sound  $IGA_{opt}^i$  was selected for each subject using the TOPSIS method.
- A probability score was calculated for each sound, based on the dependency between the design variables. The 8 sounds with the highest score were recommended.

Question Q1:

- The sounds  $IGA_{opt}^i$  obtained good performances regarding the optimization problem, comparable to a designer's proposals.

Question Q2:

- The proposed method resulted in diverse recommended sounds, with performances that were on par with a designer's proposal.
- The coverage of the Pareto front by the recommended sounds could be improved.

Figure 19. Overview of the results of Experiments 1–3.

each experiment. More specifically, the correlation between  $L_{Aeq}$  and the two objectives was calculated, based on the raw evaluations of the subjects, after eliminating outliers. The computation of  $L_{Aeq}$  is provided in the following equation:

$$L_{Aeq} = 10 \log \left( \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \frac{p_A^2}{p_{ref}^2} \right), \tag{19}$$

where  $t_1$  and  $t_2$  are, respectively, the beginning and the end of the signal,  $p_A$  is the instantaneous A-weighted sound pressure and  $p_{ref}$  is the reference sound pressure of  $20 \times 10^{-6}$  Pascals.

Table 7 shows the Pearson's correlation coefficients  $\rho$  between the subjects' evaluations and  $L_{Aeq}$  for all three experiments. It shows a significant positive correlation between unpleasantness and  $L_{Aeq}$ , as well as a significant negative correlation between detection time and  $L_{Aeq}$ . These results confirm a general trend that, on average, the louder a sound is, the faster it is to detect and the more unpleasant it becomes, which is expected.

However, the sound level is not a relevant explicating factor. For example, the  $L_{Aeq}$  is equal to 77 dB for *rand2* and 75 dB for *reco2*, but *rand2* is significantly less unpleasant (Figure 13) and significantly less detectable (Figure 14) than *reco2*. To better explain the differences between the sounds, spectral information should be

**Table 7.** Pearson's correlation coefficients between  $L_{Aeq}$  and the two objectives, computed from all subjects' evaluations, without any pre-processing. All coefficients are statistically significant ( $p < 0.001$ ).

	Experiment 1	Experiment 2	Experiment 3
Number of evaluations	2772	672	1224
Unpleasantness	$\rho = 0.286$	$\rho = 0.192$	$\rho = 0.245$
Detection time	$\rho = -0.290$	$\rho = -0.240$	$\rho = -0.296$

considered. The sound level alone would be insufficient to make design recommendations.

## 8.2. QV sound design using an IGA

The first experiment showed that an IGA can be used to interactively explore a sound design space and to define efficient solutions for a bi-objective optimisation problem. This is consistent with previous studies showing the application of IGAs as interactive design tools to address complex perceptual dimensions or affordances (Brintrup *et al.* 2008; Mata *et al.* 2018). In particular, when the two dimensions are conflicting, an IGA may provide efficient designs because interactions between variables are integrated during the interactive solutions search. The linear models without interactions fitted to the data showed opposite effects of the factors on the two objectives, making them ineffective for delivering relevant recommendations.

We believe that an IGA is not intended to replace a designer. Rather, the method could be seen as a complementary tool to assist the work of a sound designer, when the perceptual dimensions are complex or even conflicting. A designer can, for example, perform Experiment 1, in order to get a set of Pareto optimal solutions. These solutions could be considered as 'presets' of the sound synthesis method, that could be next 'fine-tuned' by the designer, using additional constraints. For QV sounds, constraints related to the brand identity may be, for example, particularly important. This idea is similar to the concept of 'perceptually relevant presets' of a musical synthesiser, presented in Roche (2020). The diversity of the proposals on the Pareto front is also a strong point of the method, in order for the designer to propose several prototypes to stakeholders. In a customisation perspective of the QV sound, it could also be interesting to present to customers via a Web interface a way to customise the sound of their vehicle using IGA. In this study, we did not take into account legal or technical considerations regarding the design. An important advantage of the proposed approach is that those constraints can easily be included in the optimisation process, either directly in the sound synthesis method, or as optimisation constraints.

## 8.3. Design recommendation method

### *Recommended sound features*

The recommendation method presented allows the uncovering of combinations of the levels of the variables that are efficient for the objectives considered in the study.

According to our results, there are four combinations of the variables (A, B, C, D) and two combinations of (E, F) that are recommended (see [Table 6](#)):

- (i) A3 B2 C2 D2;
- (ii) A1 B4 C4 D4;
- (iii) A2 B4 C4 D4;
- (iv) A1 B2 C4 D2;
- (v) A4 B3 C4 D2;
- (vi) E1 F1 (no modulation);
- (vii) E3 F2 (modulation 5 Hz – 17%).

Three pairs of sounds (*reco1, reco4*), (*reco3, reco6*) and (*reco2, reco5*) are pairwise identical on variables (A, B, C, D), and the only difference is that the second sound of the pair possesses an amplitude modulation (E3 F2). The perception of each pair is similar, the sounds being close on the plane of the two objectives ([Figure 16](#)), meaning that the effect of the variables (E, F) on the objectives is weak.

However, all the recommended designs are combinations of variable levels that affect the objectives in opposite ways: none of the recommended designs have a value of 1 or 4 for all variables and some designs have variable values that go against the linear model's prediction. For example, considering the resemblance to an ICE vehicle, *reco2* has only engine-like sound (A1 and C4), but is very unpleasant, whereas *reco7* has no engine-like component (A4), but is much less unpleasant. This illustrates that there are interactions between the design variables, and that in order to minimise both objectives, design rules have to consider groups of variables simultaneously.

A future work would be to analyse the recommended sounds according to features of the signal, in order to extract design rules. Indeed, the recommended sounds are diverse at the design variable level, and it would be useful to know what characterises them acoustically, apart from the sound level.

### **Benefits and limitations of the method**

The recommended sounds represent a diverse set of solutions to the design problem. Designers may take them into account, and tune them further according to personal preferences and other constraints or objectives that are not taken into account in the objective functions optimised. These sounds are relevant examples with varied but good performances. While the sounds from *Optimalset* are good solutions from each subject's point of view, the recommended sounds account for the perception of a whole subject group. This is particularly relevant for real-world problems, where a product's sound design has to fulfil its purpose for as many people as possible, in the target audience.

The fact that the two designed sounds were not recommended shows that even though the recommendation method is able to find some interesting sounds, it is unable to find all of them. It can be due either to the recommendation method, or to the IGA experiment. This is a limit of the method, which could not capture all the relevant trade-offs for the problem.

The efficiency of the recommendation method depends on the number of occurrences of the same designs in the middle area of the Pareto front. If there is a disagreement between the subjects, the recommendation method is unable to capture these multiple designs. This could explain why the recommended sounds tend to be located at the extremes of the Pareto front. The method assumes that



there are some acoustic features that contribute to the perception of unpleasantness and detectability, in a similar way across individuals. Studies on this topic (Misdariis *et al.* 2013; Lee *et al.* 2017) suggest that such features exist. If such features did not exist, the IGA could still be used for customisation, but there would be nothing to capture on the group level for the recommendation method. The results of Experiment 3 suggest that there is more inter-individual variability in the perception of what constitutes a good compromise between the objectives, than there is regarding what makes a sound extreme according to these same objectives.

Another explanation that can be proposed is that the IGA tends to find solutions that are dense at the extremes of the Pareto front, while being sparser in the middle. Further studies are needed to confirm this intuition. Nevertheless, the recommendations remain relevant, as some of them are on par with the second proposal of the designer, the sound *design2*.

### ***Future work and applications***

A first obvious application of the proposed method would be for the design of other sounds with similar conflicting objectives, such as alarm or mobile phone sounds.

A second important application of the recommendation method would be to investigate perceptual dimensions that are very specific, related to an atypical panel of users (amateur of very sporty cars, blind people etc.). In this case, the designer could use the recommendations based on assessments of this panel to develop prototypes or archetypal sounds. The development of sounds for very sporty cars would be, for example, a relevant application use case of the proposed method.

More generally, the proposed method applies to any problem dealing with the optimisation of a parametrised design, according to one or several perceptual quantities, as long as the evaluation task is simple and short enough to allow for interactive optimisation. Thus, it is not limited to sound design, but can also apply to product shape design (e.g., Mata *et al.* 2018). However, if the parameters take continuous values, the design recommendation method should be adapted (by, e.g., estimating probability density functions).

As it was done in this study, an IGA can also be used to study the dependencies between the design variables. When trying to build a model through a DOE approach, this information could be used to make an informed choice of the interaction terms to consider.

## **9. Conclusion**

In this paper, we presented a computational method to perform design recommendations based on an experiment with an IGA. The application is the bi-objective optimisation of the unpleasantness and the detectability of sounds for QV, and uses listening tests with a panel of participants and an IGA for the multiobjective optimisation. The results showed that (i) the individual sounds defined by the IGA perform well regarding the optimisation problem when compared with other QV sounds proposals and (ii) the design recommendations allow the definition of a large diversity of sounds. Those sounds are found to be well performing on average when compared with other proposals. In particular, some of the recommended sounds obtained very good performances in terms of detectability.

The proposed design method is rather generic, as other characteristics than detectability and unpleasantness can be optimised. For example, the sportiness of QV sounds could be investigated, as it presents a challenge for manufacturer because of the lack of background knowledge on this topic. The agreement with the brand image could also be an interesting perceptual dimension.

In light of those results, several research perspectives can be considered. The first one will be to investigate the relationship between signal parameters and the perception of the sounds, in order to extract design rules. The second perspective is the study of more specific perceptual dimensions, related to expert listeners or disabled people. The third perspective will be to guide the parameterisation of the sound synthesis, with considerations for the perceptions of listeners, in order to ensure noticeable differences between sounds available during perceptual IGA experiments.

## Acknowledgments

Acknowledgments to the students and professors of the École Centrale de Nantes for their participation to the listening tests.

## Financial support

This study was co-financed by the Pays de la Loire region (RFI Ouest Industrie Créative), France, and the European Regional Development Fund.

## References

- Bi, Y., Li, S., Wagner, D. & Reid, T. 2017 The impact of vehicle silhouettes on perceptions of car environmental friendliness and safety in 2009 and 2016: a comparative study. *Design Science* 3, E23; doi:[10.1017/dsj.2017.22](https://doi.org/10.1017/dsj.2017.22).
- Biles, J. A. 2007 Evolutionary computation for musical tasks. In *Evolutionary Computer Music*, pp. 28–51. Springer.
- Brintrup, A. M., Ramsden, J., Takagi, H. & Tiwari, A. 2008 Ergonomic chair design by fusing qualitative and quantitative criteria using interactive genetic algorithms. *IEEE Transactions on Evolutionary Computation* 12 (3), 343–354.
- Carron, M., Rotureau, T., Dubois, F., Misdariis, N. & Susini, P. 2017 Speaking about sounds: a tool for communication on sound features. *Journal of Design Research* 15 (2), 85–109.
- Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. 2002 A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6 (2), 182–197.
- Delle Monache, S., Rocchesso, D., Bevilacqua, F., Lemaitre, G., Baldan, S. & Cera, A. 2018 Embodied sound design. *International Journal of Human-Computer Studies* 118, 47–59.
- Desoeuvre, G., Richard, F., Roussarie, V. & Bezat, M.-C. 2008 Hartis, a re-synthesis tool for vehicles sound design. *The Journal of the Acoustical Society of America*. 123 (5), 3247–3247.
- Durant, E. A., Wakefield, G. H., Tasell, D. J. V. & Rickert, M. E. 2004 Efficient perceptual tuning of hearing aids with genetic algorithms. *IEEE Transactions on Speech and Audio Processing* 12 (2), 144–155.

- Edworthy, J., Loxley, S. & Dennis, I.** 1991 Improving auditory warning design: relationship between warning sound parameters and perceived urgency. *Human Factors* **33** (2), 205–231.
- Engler, O.** 2016 Mercedes-Benz SLS AMG coupé electric drive: NVH development and sound design of an electric sports car. SAE Paper No. 2016-01-1783.
- Eppinger, S. D. & Ulrich, K. T.** 2016 *Product Design and Development*. McGraw-Hill Education.
- Fagerlönn, J., Larsson, P. & Maculewicz, J.** 2020 The sound of trust: sonification of car intentions and perception in a context of autonomous drive. *International Journal of Human Factors and Ergonomics* **7** (4), 343–358; doi:[10.1504/IJHFE.2020.112506](https://doi.org/10.1504/IJHFE.2020.112506).
- Goldberg, D. E.** 1989 *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley.
- Gu, Z., Tang, M. X. & Frazer, J. H.** 2006 Capturing aesthetics intention during interactive evolution. *Computer-Aided Design* **38**, 224–237.
- Hwang, C.-L. & Yoon, K.** 1981 *Multiple Attribute Decision Making*. Lecture Notes in Economics and Mathematical Systems, 186. Springer; doi:[10.1007/978-3-642-48318-9](https://doi.org/10.1007/978-3-642-48318-9).
- Ibrahim, A. A. A., Yiap, S. Y. & Andrias, R. M.** 2018 Kansei engineering concept in sound design. *Advanced Science Letters*. **24**, 1320–1324; doi:[10.1166/asl.2018.10741](https://doi.org/10.1166/asl.2018.10741).
- Japan Automotive Standards Internationalization Center** 2009 A study on approach warning systems for hybrid vehicle in motor mode. Informal document No. GRB-49-10 (49th GRB).
- Kerber, S. & Fastl, H.** 2008 Prediction of perceptibility of vehicle exterior noise in background noise. In *Proceedings of DAGA 2008, Dresden, October 2008*, pp. 623–624. DEGA.
- Khuri, A. I., Mathew, T. & Sinha, B. K.** 1998 *Statistical Tests for Mixed Linear Models*. John Wiley.
- Konet, H., Sato, M., Schiller, T., Christensen, A., Tabata, T. & Kanuma, T.** 2011 Development of approaching vehicle sound for pedestrians (VSP) for quiet electric vehicles. *SAE International Journal of Engines* **4** (1), 1217–1224.
- Lee, S. K., Lee, S. M., Shin, T. & Han, M.** 2017 Objective evaluation of the sound quality of the warning sound of electric vehicles with a consideration of the masking effect: annoyance and detectability. *International Journal of Automotive Technology* **18** (4), 699–705.
- Maffiolo, V.** 1999. *De la caractérisation sémantique et acoustique de la qualité sonore de l'environnement urbain: structuration des représentations mentales et influence sur l'appréciation qualitative: application aux ambiances sonores de Paris*. PhD Dissertation, Le Mans.
- Marshall, D. C., Lee, J. D. & Austria, P. A.** 2007 Alerts for in-vehicle information systems: annoyance, urgency, and appropriateness. *Human Factors* **49** (1), 145–157.
- Mata, I., Fadel, G., Garland, A. & Zanker, W.** 2018 Affordance based interactive genetic algorithm (ABIGA). *Design Science* **4**, E5; doi:[10.1017/dsj.2017.30](https://doi.org/10.1017/dsj.2017.30).
- Miki, M., Orita, H., Wake, S. H. & Hiroyasu, T.** 2006 Design of sign sounds using an interactive genetic algorithm. In *2006 IEEE International Conference on Systems, Man and Cybernetics, Taipei*, pp. 3486–3490. IEEE.
- Misdariis, N., Cera, A., Levallois, E. & Locqueteau, C.** 2012 Do electric cars have to make noise? An emblematic opportunity for designing sounds and soundscapes. In *Proceedings of the Acoustics 2012 Nantes Conference*, Nantes, France, pp. 1045–1050. Acoustics 2012.

- Misdariis, N., Cera, A. & Rodriguez, W. 2019 *Electric and Autonomous Vehicle: From Sound Quality to Innovative Sound Design*. International Congress on Acoustics. <https://hal.archives-ouvertes.fr/hal-02469333>.
- Misdariis, N., Gruson, A. & Susini, P. 2013 Detectability study of warning signals in urban background noises: a first step for designing the sound of electric vehicles. *Proceedings of Meetings on Acoustics* **19**, 040032; doi:10.1121/1.4799454.
- Nagamachi, M. 1995 Kansei engineering: a new ergonomic consumer-oriented technology for product development. *International Journal of Industrial Ergonomics* **15**, 3–11.
- Orsborn, S., Cagan, J. & Boatwright, P. 2009 Quantifying aesthetic form preference in a utility function. *Journal of Mechanical Design* **131** (6), 061001.
- Özcan, E. & Van Egmond, R. 2012 Basic semantics of product sounds. *International Journal of Design* **6** (2), 41–54.
- Parizet, E., Ellermeier, W. & Robart, R. 2014 Auditory warnings for electric vehicles: detectability in normal-vision and visually-impaired listeners. *Applied Acoustics* **86**, 50–58.
- Parizet, E., Robart, R., Chamard, J.-C., Schlittenlacher, J., Pondrom, P., Ellermeier, W., Biancardi, F., Janssens, K., Speed-Andrews, P., Cockram, J. & Hatton, G. 2013. Detectability and annoyance of warning sounds for electric vehicles. *Proceedings of Meetings on Acoustics* **19**, 040033; doi:10.1121/1.4800182.
- Pedersen, T. H., Gadegaard, T., Kjems, K. & Skov, U. 2011 *White Paper on External Warning Sounds for Electric Cars-Recommendations and Guidelines*. Delta.
- Petiot, J.-F., Kristensen, B. G. & Maier, A. M. 2013 How should an electric vehicle sound? User and expert perception. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 55928)*, p. V005T06A028. American Society of Mechanical Engineers.
- Petiot, J.-F., Legeay, K. & Lagrange, M. 2019 Optimization of the sound of electric vehicles according to unpleasantness and detectability. *Proceedings of the Design Society: International Conference on Engineering Design* **1** (1), 3949–3958.
- Petiot, J.-F., Villa, S., Denjean, S. & Diaz, E. (2020). Design of warning sounds using an interactive genetic algorithm. In *Proceedings of Forum Acusticum 2020*, December 2020, Lyon, France, pp. 107–114. Forum Acusticum 2020.
- Poirson, E., Petiot, J. F., Boivin, L. & Blumenthal, D. 2013 Eliciting user perceptions using assessment tests based on an interactive genetic algorithm. *Journal of Mechanical Design, Transactions of the ASME* **135** (3), 031004; doi:10.1115/1.4023282.
- Poveda-Martínez, P., Peral-Orts, R., Campillo-Davo, N., Nescolarde-Selva, J., Lloret-Climent, M. & Ramis-Soriano, J. 2017 Study of the effectiveness of electric vehicle warning sounds depending on the urban environment. *Applied Acoustics* **116**, 317–328.
- Pulkki, V. 2001 Localization of amplitude-panned virtual sources II: two- and three-dimensional panning. *Journal of the Audio Engineering Society* **49**, 753–767.
- Roads, C. 1996 *The Computer Music Tutorial*. MIT Press.
- Robart, R. & Parizet, E. 2013. eVADER: a perceptual approach to finding minimum warning sound requirements for quiet cars. In *AIA-DAGA 2013 Conference on Acoustics*, p. 1. DEGA.
- Roche, F. 2020 Music sound synthesis using machine learning: towards a perceptually relevant control space. Thèse de l'Université Grenoble Alpes, September 2020. HAL Id: tel-03102796, version 1.
- Singh, S., Payne, S. R. & Jennings, P. A. 2014 Toward a methodology for assessing electric vehicle exterior sounds. *IEEE Transactions on Intelligent Transportation Systems* **15** (4), 1790–1800.

- Souaille, T., Petiot, J.-F., Lagrange, M. & Misdariis, N.** 2021 Extracting design recommendations from interactive genetic algorithm experiments: application to the design of sounds for electric vehicles. In *Proceedings of the International Conference on Engineering Design (ICED21)*, Gothenburg, Sweden, 16–20 August 2021; doi:[10.1017/pds.2021.418](https://doi.org/10.1017/pds.2021.418).
- Steinbach, L. & Altinsoy, M.** 2019 Prediction of annoyance evaluations of electric vehicle noise by using artificial neural networks. *Applied Acoustics* **145**, 149–158.
- Swart, D. J., Bekker, A. & Bienert, J.** 2018 The subjective dimensions of sound quality of standard production electric vehicles. *Applied Acoustics* **129**, 354–364.
- Sylcott, B., Michalek, J. J. & Cagan, J.** 2015 Exploring the role of interaction effects in visual conjoint analysis. *Journal of Mechanical Design* **137** (9), 094503.
- Takagi, H.** 2001 Interactive evolutionary computation: fusion of the capabilities of EC optimization and human evaluation. *Proceedings of the IEEE* **89** (9), 1275–1296.
- Teichert, T. & Shehu, E.** 2007 Evolutionary conjoint. In *Conjoint Measurement*, pp. 113–132. Springer.
- Wakefield, G. H., Van Den Honert, C., Parkinson, W. & Lineaweaver, S.** 2005 Genetic algorithms for adaptive psychophysical procedures: recipient-directed design of speech-processor MAPs. *Ear and Hearing* **26**, 57S–72S.