Sound quality assessment: comparison of in-situ and on-line experiments

Etienne Parizet[1], Dorian Grappe[1], Chaouki Benzekri[1], Clément Coppel[2]

[1]Univ. Lyon, INSA-Lyon, Laboratoire Vibrations Acoustique, Villeurbanne, France
[2]Volvo Group Trucks Technology, Saint-Priest, France
etienne.parizet@insa-lyon.fr

Abstract

In order to assess the sound quality of electrically powered trucks, a listening test experiment was planned in 2020. 27 sounds recorded in different vehicles were equalized in loudness (using the Zwicker model) to highlight the contribution of timbre parameters. Due to the health situation and the lockdown in France, this experiment was conducted online; 55 people participated. The following year, it was repeated in the laboratory and at the truck manufacturer (46 participants). In the online experiment, there was no control over the experimental conditions (background noise in the room, headphones, etc.). In contrast, the in-situ experiment was carried out in soundproof booths with high-quality headphones. The objective of this presentation will be to present the comparison of the results obtained in the two conditions. Despite the lack of control in the in-situ conditions, the results of the two experiments are very close - showing that online experiments can be used for sound quality applications.

Keywords: Sound quality evaluation, on-line experiment.

1 Introduction

Most of the time, perceptual experiments are conducted in a laboratory, in a very controlled environment: participants are in a room isolated from outside noise, they are focused on their task and sound are presented with high-fidelity headphones or loudspeakers. Very few experiments are conducted via the Internet (an example is given in [1]). Indeed, the conditions of such an experiment are not controlled: does a listener use headphones or a loudspeaker to listen to the stimuli? (see [2] for a proposed protocol to check this). What is the quality of the playback system? Is the sound environment quiet enough not to disturb the listener? All this can greatly reduce the quality of the results.

Because the Covid pandemic led to a lockdown in 2020, followed by a period during which outsiders could not enter the laboratory, an online experiment had to be set up. We wandered whether the results of the experiment could be considered as reliable. To get an answer, this experiment was repeated in 2021 under usual controlled conditions.

2 Methods

2.1 Stimuli

These experiments were conducted as part of a study about noise comfort in electrically powered trucks. The objective of the study was to identify the sound timbre parameters that contribute to the acoustic quality of the vehicles. The sounds of several vehicles were recorded (f_s=51200 Hz) at the driver's ears, for different speeds (from 30 to 90 km/h). In some stimuli, emerging frequencies (between 600 and 3000 Hz) were amplified (by 6 to 15 dB). Thus, 27 stimuli were obtained – they had a duration of 10 s. Fading in and out was applied (250 ms duration).
The large differences in vehicle speeds resulted in large differences in level (about 20 dBA between the minimum and maximum values). As perceived loudness is a very important factor in acoustic quality, it was decided to equalise this loudness. An iterative procedure was used, consisting in equalizing calculated according to ISO 532-1 (2017). The metric was the average of loudness computed at both ears, equalized to a level of about 19 soneGF (with a total deviation between all signals of less than 0.5 soneGF).

2.2 On-line experiment (2020)

As the health situation prohibited participants to come to the laboratory, it was decided to carry out an on-line experiment using Psytoolkit [3,4]. After listening to each sound, the participant had to evaluate four parameters: *loudness*, *sharpness*, *regularity* and finally *pleasantness*. For each descriptor, the response was to be given by moving a slider on a continuous scale with five extreme labels (e.g. from *extremely unpleasant* to *extremely pleasant*). The position of the cursor was recorded as an integer between 0 and 100.
Beforehand, five signals were presented to allow the listener to appreciate the variation of the parameters. In addition, the participant was asked to indicate his or her age (in four categories: *18-25 years*, *26-40 years*, *41-60 years* and *over 60*), gender and experience of driving a truck (*never* / *rarely* / *sometimes* / *regularly*). Finally, as it was not possible to calibrate the listening level, the participant was asked to listen to a speech signal and to adjust the level of the computer so that the speech was just intelligible. A limitation of Psytoolkit is that it is not possible to combine two elements:
- a random presentation of signals;
- the presentation of a progress indicator for each sound (which is useful to the participant).

We therefore divided the experiment into four blocks (presented randomly and within which the order of the sounds was also random), with a progress indicator presented to the participant at the end of each block.
The link was disseminated and this allowed to collect 55 complete responses within a couple of months (other participants dropped out along the way).

2.3 Laboratory experiment (2021)

The experiment was repeated in 2021. This time, participants were allowed to come to the laboratory. Each participant was placed in an sound-proof booth and sounds were presented through headphones (Sennheiser HD650) at a calibrated level corresponding to an average loudness of 19 soneGF. 46 people took part in the experiment: 30 students from INSA-Lyon (aged 18 to 25 and with no experience of driving industrial vehicles) and 16 members of the manufacturer's staff, of various ages and all with driving experience.
The rest of the procedure was identical to the experiment conducted in 2020, with the exception that the speech signal was not presented (as participants were not allowed to modify the presentation level of the sounds). This experiment was also run by Psytoolkit, so that only the listening and environmental conditions differed between the two experiments.

3 Results

For each descriptor, an analysis of variance (repeated measures) was conducted. It included a within-subjects factor (stimuli, 27 levels) and a between-subjects factor (type of experiment: on-line or in-situ). These analyses were conducted by JASP (https://jasp-stats.org). The Greenhouse-Geisser sphericity correction was used and the significance threshold was 5%.

For the parameters *Sharpness* and *Regularity*, the results are similar between the two experiments ($p>0.05$). A difference is noted for the parameters *Loud* ($F(26,2574)=2.48$, $p=0.003$), and *Pleasant* ($F(26,2574)=1.59$, $p=0.021$), as shown in Figures 1 and 2.

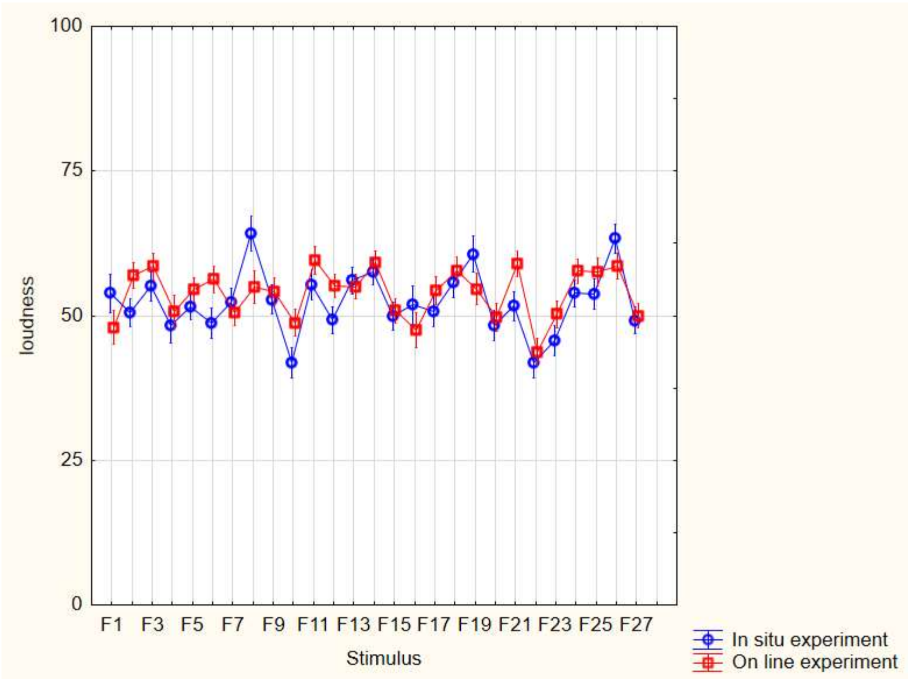

Figure 1 – Averaged loudness of the 27 stimuli. Blue circles : on-line experiment (2020). Red squares : in situ experiment (2021).

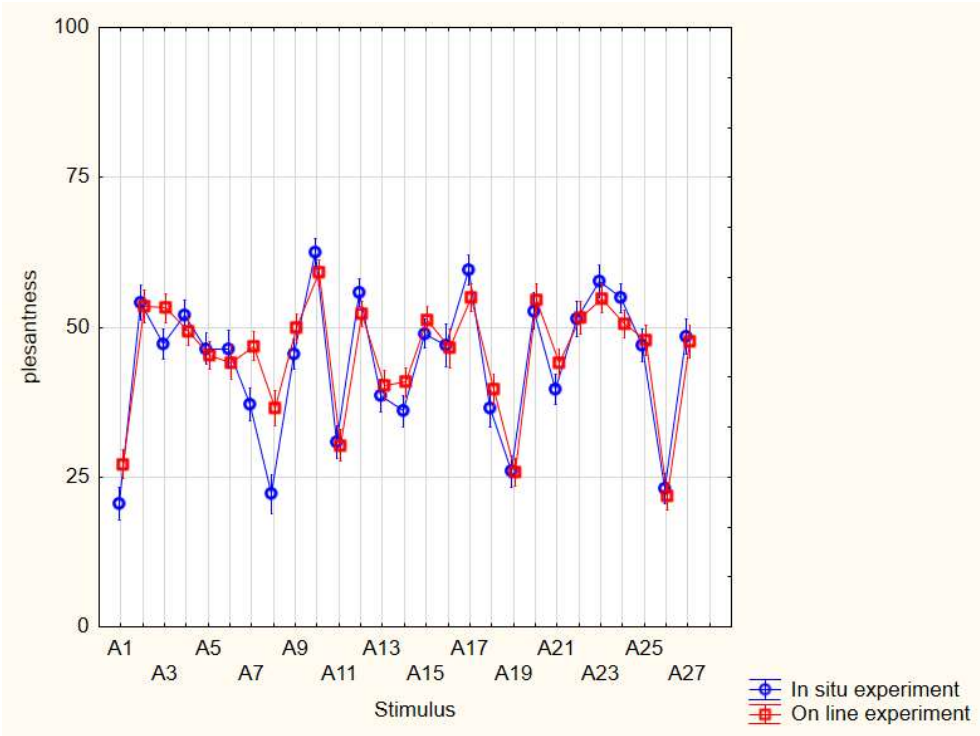

Figure 2 – Averaged pleasantness of the 27 stimuli. Blue circles : on-line experiment (2020). Red squares : in situ experiment (2021).

Although the statistics relate a significant difference between the two experiments, figures 1 and 2 shows that results are quite similar. Moreover, it can be suspected that these differences are due to a different age distribution of the listeners in the two experiments. The age distribution was not similar for the two experiments (table A). Distributions are significantly different ($\chi^2(3) = 11.6, p < 0.01$).

	18 to 25	26 to 40	41 to 60	> 60
On line (2020)	19	13	21	2
In situ (2021	31	10	5	0

Table A - Age distribution of participants in the two experiments

If the analyses of variance are repeated for the younger participants only (between 18 and 25, which represents the largest category in both cases), the effect of the type of experiment is not significant for any of the four sound parameters.

4 Conclusions

The results obtained in the two experiments are very similar. It therefore seems possible to carry out online experiments more often. An advantage of this type of experiment is that, potentially, a very large number of people can participate. A limitation is that it is difficult to target a particular profile of participants (for example, in our case, regular drivers of industrial vehicles). It therefore seems reasonable to limit the use of online experiments to cases where such particular profiles are not sought.

References

[1] Luo, L., Zheng, X., Hao, Z.Y., Dai W.Q., Yang W.Y. (2017). Sound quality evaluation of high-speed train interior noise by adaptive Moore loudness algorithm. Journal of Zhejiang University-Science A 18(9), 690-703.

[2] Woods, K., Siegel, M., Traer J., McDermott J. (2017). Headphone screening to facilitate web-based auditory experiment. Attention, Perception & Psychophysics 79(7), 2064-2072.

[3] Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42(4)*, 1096-1104.

[4] Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44(1)*, 24-31.