Comparison of listening test methods : a case study

Etienne Parizet, Nacer Hamzaoui, Guillaume Sabatié

Laboratoire Vibrations Acoustique, Insa Lyon, F-69621 Villeurbanne, France, Email : parizet@lva.insa-lyon.fr

Introduction

For sound quality applications, many listening tests methods can be used, which have the same goals : to compare the pleasantness of different sounds, to identify the timbre aspects which are used by listeners to give their assessment and, finally, to build a indicator of that assessment from usual sound metrics.

But a comparison of different method has rarely been conducted [1,2,3,4]. The goal of this study was to do such a comparison, in the limited case of some procedures and a precise type of sounds.

Experiment

Stimuli

Sounds of ventilation system in four cars were used. The car was stopped with the engine off while the ventilation system was switched on with different settings (fan rotational speed, heating or air conditioning, etc). A dummy head (Bruel et Kjaer) was located on the driver's seat. Among the whole collection of data, nine sounds were selected, because they had very similar loudness (the maximum loudness difference was 1 Phone, when computed from an ISO 532B loudness software). Each signal had a duration of 10 seconds; they were presented to listeners through headphones (Sennheiser HD600), in a quiet room, at an averaged level of 74 dB(A).

Test procedures

Six test procedures were used. In the first one, the listener was presented sounds one by one. After hearing a sound, he had to give his answer on a scale going from "very unpleasant" to "very pleasant". In the second one, the same scale was used; the difference was that all scales were presented on the computer's screen. Beside each scale was a button allowing the subject to hear again the corresponding sound. Such a procedure is therefore a mix between evaluation and comparison. Then three pair comparison tests were conducted. The first one was a forced choice test (the listener had to select one of the sounds as the preferred one); in the second one, the listener had the choice between five answers, one of which being that the two sounds are felt equally pleasant, and the third test proposed a continuous scale to the listener. Finally, the sixth procedure was a similarity rating, for which the answer was also given on a continuous scale.

In the following, these tests procedures will be labelled T1 to T6.

For the first test, the order of presentation of sounds was randomized. For the last four ones (for which sounds were presented in pairs), the set of pairs were ordered according to Ross series.

64 subjects participated to the experiments in two sessions., separated by approximately a week. T1 and T2 did not belonged to the same session, so did not T4 and T5.

After each test, the listener was asked to evaluate its length and difficulty on two scales.

Results

Duration and difficulty of tests

Figure 1 presents the averaged estimated length of each test, as well as their averaged real duration, showing the clear relation between these two values.

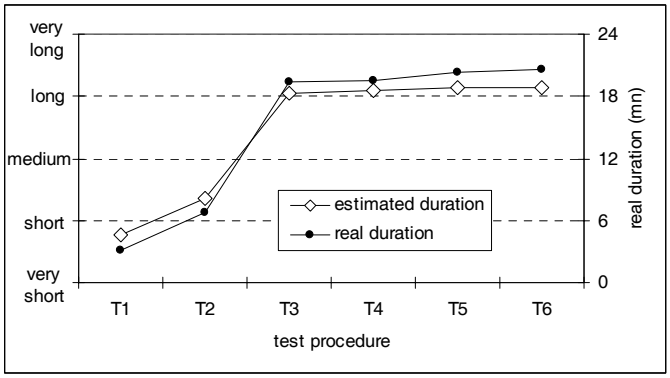

Figure 1: Estimated and real duration of each test

The estimated difficulty was quite similar within procedures. Nevertheless, it was significantly greater for T6 (similarity ratings) than for the other ones.

Merit scores of noise

The preference answers obtained in tests 3 to 5 were converted in merit scores by the simple formula

$$S_i = \sum_{j \neq i} P_{ij} \qquad (1)$$

which allowed to compare their results with those of the evaluation tests (T1 and T2). As is shown in figure 2, these merit scores are very close to each others.

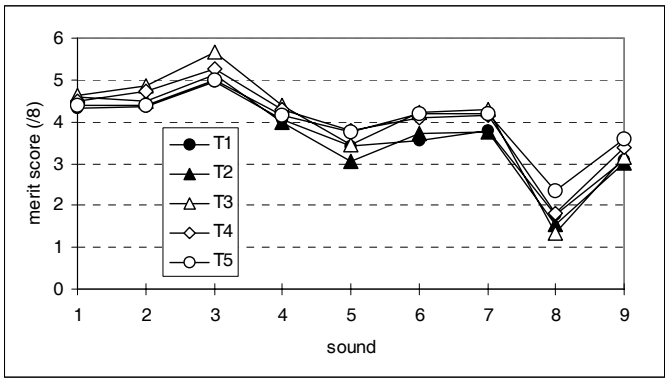

Figure 2 : merit scores of noises obtained in tests 1 to 5

But a closer look at the results of test T1 show that they are different between listeners who started their session by that test (half of the jury) and those who passed it at the end of the session (the other half), which was not the case for T2. Therefore, T1 seems to give less accurate results than T2.

Also, for each of the first five tests, a separation of the jury in homogeneous groups of listeners was realised using the K-means technique. The results were very stable among tests, as a two-classes splitting was always a good solution, and the number of listeners in each class were similar (table 1). Moreover, 34 listeners out of 64 always belonged to the same group.

Perceptive space

The results of each test was finally used to build the perceptual space of sounds. For the first five ones, a Principal Component Analysis was realised from the noises merit scores obtained from each listener, using equation (1). For the similarity evaluation (T6), an Indscal analysis, as defined by Carroll and Chang was conducted. The dimensions of the spaces thus obtained were very similar (figure 3).

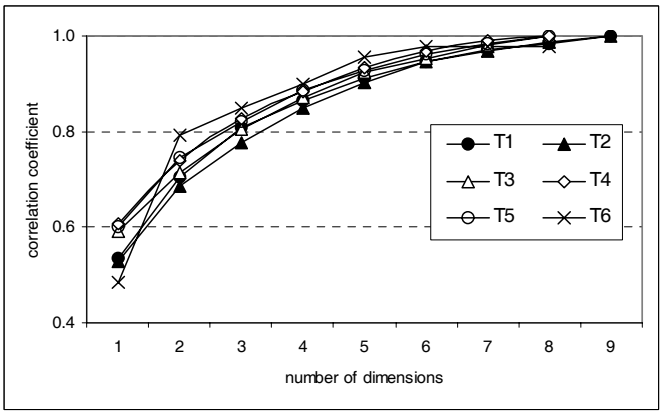

Figure 3 : cumulated variance explained in the PCA of tests 1 to 5 and in the Indscal analysis of test 6

The perceptual spaces obtained from these analysis were also similar, with the exception of test 1 which give slightly different results. The first axis of the space stemmed from the Indscal analysis of similarity evaluations could be linked to the frequency balance of sounds, as noises co-ordinates over that axis had a good correlation with their sharpness values (table 2).

T1	T2	T3	T4	T5	T6
0.44	0.82	0.95	0.92	0.93	0.99

Table 2 : correlation between sharpness values and co-ordinates of noises on the first axis of each analysis

The second axis of the perceptual space given by the Indscal analysis could also be related to a sound feature (namely, the presence of a strong tonal component in one of the noises), which could not be done for the other analysis.

It was also possible to build an accurate model of merit scores using the co-ordinates of sound on the Indscal space. Merit scores were computed from the results of test T4 (pair comparison with a five-levels scale), computed over each of the two groups of listeners. The models were :

$$\begin{cases} S^{T4}_{Group1} = 4 - 4.5X^{T6}_1 - 5.6X^{T6}_2 \\ S^{T4}_{Group2} = 4 + 4.9X^{T6}_1 - 5.0X^{T6}_2 \end{cases} \quad (2)$$

X^{T6}_1 and X^{T6}_2 being the co-ordinates of sounds on the two axis. For each group, the correlation coefficients between measured and predicted scores were very high (R=0.93). Also, equations (2) explained the differences between the two groups, related to the frequency balance of sounds (the coefficients of X^{T6}_1 have opposite signs).

On the other hand, it was not possible to obtain such a precise model from the co-ordinates of sounds on the first two axis of a Principal Component Analysis of one of the first five tests. This indicates that, in this studied case, a similarity evaluation provided more information about the perceptual space.

References

[1] "Verbal attributes of simultaneous wind instruments", Kendall R. , Carterette E., Music Perception 10(4), 1993, 445-468

[2] "Measurement of quantities depending upon perception by jury-test methods", Rossi G. et al., Measurement 34, 2003, 57-66

[3] "Psychophysical scaling of the pleasantness of everyday-noises", Zeitler & Hellbruck, Proc. 8th Oldenburg Symp (2000).

[4] "Multi-dimensional listening test : selection of sound descriptors and design of the experiment", Parizet & Nosulenko, Noise Control Eng. Journal 47(6), 1999, 1-6.