# Multi-dimensional listening test: selection of sound descriptors and design of the experiment

Etienne Parizet, Valery Nosulenko

Noise Control Eng. J. 47 (1999), 1-6

# Multi-Dimensional Listening Test: Selection of Sound Descriptors and Design of the Experiment

Etienne Parizet [(a)] and Valery N. Nosulenko[(b)]

( Received 1998 September 27; revised 1999 June 12; accepted 1999 July 08)

**A method for the selection of parameters used in multi-dimensional listening tests is presented. It allows the presentation to subjects, through a very short learning process, of a complete set of parameters which are unambiguously understood. This set is built after a first test in which other listeners are asked to freely describe differences and similarities between sounds. A method of analysing these free verbalizations allows the determination of peculiarities for sounds belonging to the same context.**

**Two multi-dimensional test methods are compared. The method in which all sounds (or pairs of sounds) are evaluated according to each parameter give slightly more reliable results than the classical method, in which each sound or pair of sounds is evaluated according to all parameters before repeating the operation for the next pair or sounds.**

**These two methods are examined using the idling noise of small diesel cars.**

Primary subject classification : 63.7; Secondary subject classification : 63.2

## 1. INTRODUCTION

Acoustic comfort in road vehicles is an important factor in customers' overall assessment of the vehicle. Car manufacturers are therefore making efforts to constantly improve this comfort. This involves obtaining a better knowledge of the acoustic perception of the various noises heard in an automobile. In particular, it is useful, for a car manufacturer to know the sound characteristics which are annoying or pleasant. This knowledge allows optimization of the acoustic design to emphasize the “good” sound parameters; it also makes it possible to create specifications to guide automobile acoustic design.

There are several methods for determining these relevant characteristics in a particular sound context. One possibility is the use of similarity tests which enable determination of a perception space with a small number of dimensions[1-2]. The experimenter then has to interpret the meaning of each dimension by identifying physical or psychoacoustic parameters correlated with the coordinates of the various stimuli on each axis. A second test, this time a preference test, then allows a preference model to be built using the parameters previously identified.

To avoid having to identify the relevant sound parameters, it is possible to use multi-dimensional methods. These methods involve evaluating sounds according to a parameter related to subjective preference (or annoyance) and a set of other more objective parameters (loud, rumbling, etc.) or subjective parameters (e.g. powerful, in the case of automobiles). There exist numerous variants of these methods: test by semantic differences, in which a sound has to be located on a set of scales opposing two terms (loud/soft, powerful/weak, etc.)[3]

[(a)] Renault, 67 rue des bons raisins, 92508 Rueil-Malmaison, France

[(b)] Anvie, 54 boulevard Raspail, 75006 Paris, France or paired comparisons according to the various parameters ("which is the loudest sound?", etc.). Such methods have been used for various vehicle noises from automobile engines[4-6] and from the noise of tires on the highway[7]. These methods present two major risks, however. First, the experimenter may forget a parameter important for perception; this can be prevented by oversizing all the parameters presented to the listener, which leads to long and tedious tests. It is also difficult to ensure that all the listeners will give the same meaning to each parameter. Semantic ambiguities may persist which produce unwanted variability in the experimental results. Of course, it is theoretically possible to train the listeners in order to standardize the understanding of the terms (this is the approach used by sensory analysis). However, it is sometimes difficult to create two sounds which are only different by a single parameter, especialy when this parameter is a complex one (for example, powerful). Moreover, often one wants to have the opinion of non-expert listeners, which is in contradiction to an extensive learning process.

The purpose of this study, therefore, therefore to outline a method by which, for a given body of sounds, multi-dimensional tests can be performed by presenting to the listeners a minimum set of parameters, understood unambiguously by all subjects and incorporating most of the sound aspects important for an overall assessment. This method was applied in an industrial context using the idling noise of diesel engines in passenger cars.

Another goal of this study was to compare two ways of carrying out a multi-dimensional listening test. In the first approach (the most frequently used), the listener, when hearing a sound or a pair of sounds, must evaluate it according to the whole set of parameters before listening to the next stimulus. The second approach consists of asking the listener to evaluate all of the sounds (or pairs of sounds) according to the first parameter. When this is completed, the subject has to evaluate the same sounds according to the next parameter, and so on. Is one of these two methods more accurate or simpler than the other? One of the aims of this study was to answer this question.

## 2. CONSTRUCTION OF THE PARAMETER BASE

### 2.A. Method

The method has been carefully described in a previous reference paper[8]. The subject hears pairs of sounds for which he has to estimate the similarity and compares the annoyance or the pleasure, justifying his choices freely. The basic principles are as follows:

- verbal comments made by subjects during a perceptual or cognitive activity are relevant indicators of this activity and can be considered as representative data for its study.
- the task of comparison imposed on the subject is a systemic factor in the perceptual, cognitive and oral communication processes;
- the comparison task can be analysed according to various dimensions: logical, perceptual and semantic.

The responses of the listeners are recorded on a tape recorder, then analysed by a very strict method [8-10]. The various verbal units are described by a series of parameters, which can then be used to create classes of equivalent verbal units. We thus obtain a refined description of all the sound aspects which have been described by the listeners, hence used by them to evaluate the annoyance.

## 2.B. Experiment

The internal noise of seven small-size vehicles with diesel engines rotating at idling speed were recorded using an acoustic dummy head. Five-second samples were built. Then 21 pairs were built (this number corresponds to a 7x7 half-matrix without the diagonal), which were stored on a digital audio tape recorder (DAT). These pairs were presented to the listeners via electrostatic headphones in a quiet room.

Each subject first heard the seven noises, then three training pairs, and finally the experimental 21 pairs. The subjects could listen to each pair again as often as they he wanted. They then performed the following three tasks: (1) give a numeric assessment of the dissimilarity of the two sounds on a scale of 0 to 8, (2) select the preferred sound, and finally (3) describe verbally the similarities and differences between the sounds, explaining the reasons for the choice.

The complete test lasted between 20 and 40 minutes. After completing the task for any pair, a listener could make a small pause to rest.

Seventy-two subjects took part in the test (51 men and 21 women). About twenty are Renault employees (about ten being noise engine specialists), and about fifty are customers driving in similar type cars.

## 2.C. Results

Analysis of the verbal comments made it possible to define seven families of equivalent parameters. Some families are not surprising:

- "Pleasant": since listeners had to compare the pleasure of the sounds, it is natural to find in their descriptions terms related to this pleasure (e.g., "A is more annoying", "B is unpleasant", etc.).
- "Loud": sound level is often the first cause of annoyance, for engine noise[11-12] as for other noises, like road noise[7]. This appear here through the comments "A is softer in level", "B is more intense", etc.
- "Sharp": a diesel engine noise can be more or less high-pitched, depending on the specific characteristics of the engine or and sound filtering by the vehicle body. The terms "Deep" and "Sharp" were therefore used frequently.

In addition to these three families, there were four other less obvious families which, for reasons of confidentiality, cannot be specified here.

For each parameter, its relevance in the characterization of each sound was determined by computing a set of values ($F_i$), in which $F_i$ is proportional to the number of occurrences of this parameter in the verbalizations related to the sound i (in the six pairs in which that sound was present). The method of computing these values is described in a previous report[9]. An example is shown in Figure 1, for the “Loud” parameter. It can be seen that the sound of car “G” is very loud, while car “E” is the most quiet one.

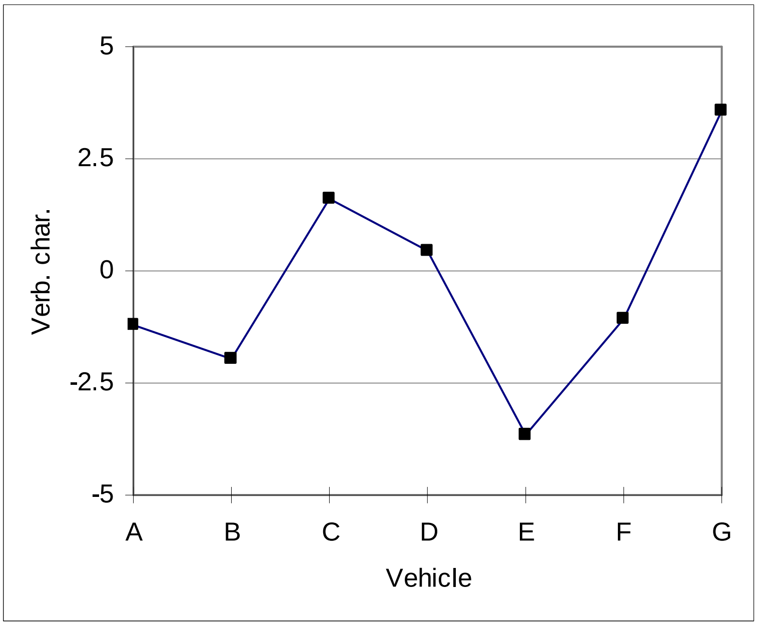


**Figure 1** : Loudness characterisation of the seven noises.

Therefore, a set of comparisons between noises was obtained. Each noise can be characterised by its most prominent features.

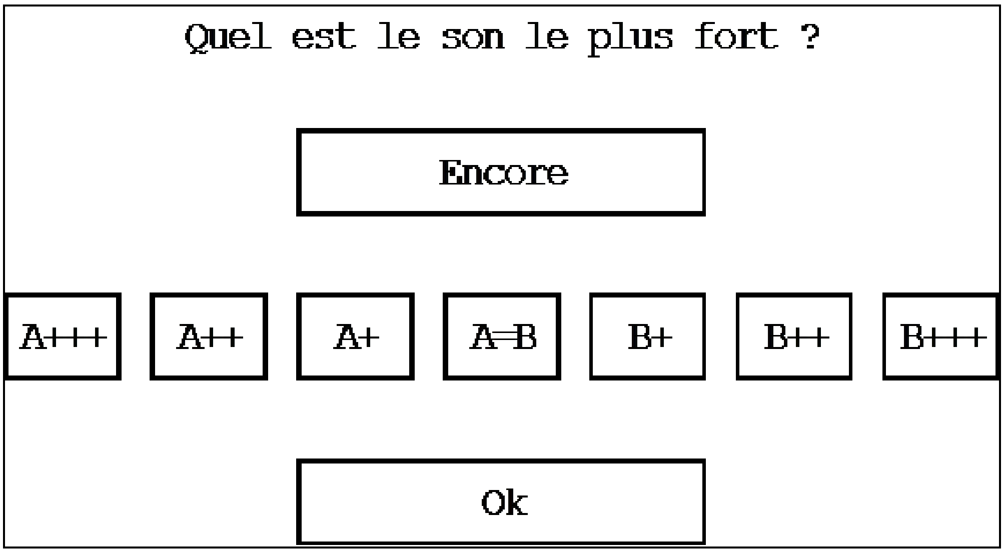
## 3. DESCRIPTION OF MULTI-DIMENSIONAL TESTS

The seven parameters above were used to build a multi-dimensional test in which, to go faster, only five noises were used (namely, cars labelled B, C, D, E and G in Figure 1). The objective was to check that, with a very short learning process which will be explained later on, these parameters can be used in a multi-dimensional test without any ambiguity. If that new test leads to the same noise description as the previous one, that would mean that the parameters are clear enough and have the same meaning for all the subjects.

### 3.A. Procedures

As another goal of the study was to compare two different ways of conducting a multi-dimensional test. Two tests were carried out.

The first one used the conventional procedure by which, for each pair, the two noises are compared according to the set of parameters adopted. Here, the entire test was controlled by a desktop computer. The various pairs appeared in random order and were delivered to the listener by the same audio set as in the first experiment. The computer presented to the listener the various parameters, likewise in random order. For each parameter there appeared a particular question (e.g. "which is the loudest noise?"), to which the listener had to reply by clicking, with the computer mouse, a box on a 7-level scale ("A much louder", "A louder ", "A slightly louder", "A and B equally loud", etc.). The subject could listen to the pair again as often as necessary, by clicking a "Once more" box. When the subject had validated their reply by clicking an "OK" box, the computer displayed the question concerning the next parameter. An example of a screen as seen by the listeners is shown in Figure 2. When all the parameters had been examined, the test moved on to the following pair. At the start of the test, the subject heard all five noises, then had to reply to a training pair.



Quel est le son le plus fort ?

**Figure 2** : Example of the screen of the computer presented to the subjects of the multi-dimensional test (“Quel est le son le plus fort ?” : “Which sound is the loudest ?” and “Encore” : ”Once more”).

The answers were stored on the computer's hard disk in the form of numbers ranging between -3 and +3, with the number of repeat hearings of the pair for each parameter and the time required to answer. Specially developed software then analysed the responses to put the pairs and the descriptors in order again.

The second test adopted the reverse procedure. The subject was requested to evaluate all the pairs according to a randomly chosen parameter. Then the same was to be done with a second parameter, and so on. For each parameter, the series of pairs were preceded by a training pair. The choice presentation, the entry of replies and the data analysis were similar to those for the preceding test.

At the end of each test, the subject had to answer a rapid questionnaire to evaluate the length and difficulty of the test. Twenty-one listeners took part in these tests (10 for the first and 11 for the second, including four women in each case). They were Renault employees, specialists neither in acoustics nor in engines.

To eliminate any risk of semantic ambiguity, the subjects underwent a very short learning process as follows: in the first procedure, the subject was shown a table in which were reported, for each parameter, some excerpts of verbal comments taken from the preceding experiment, related to the current parameter or its opposite. For example, for the "loud" parameter, this table reported the following statements : "this soud is loud", "the level is low". This was sufficient because the "loud" aspect is sufficiently common; for less obvious parameters, at most 6 excerpts were presented. This table remained before the subject’s eyes throughout the test.
In the second procedure, for each parameter, only the excerpts related to that parameter were shown to the subject.

## 4. RESULTS AND DISCUSSION

### 4.A. Comparison of the two multi-dimensional test procedures

#### a. Comparison of tests results

Fore each parameter, the average answers given by the two test procedures were computed. These average answers are not statistically different for any pair of parameter (at the 5% confident level). The correlation coeeficient between thes means are presented in Table 1; results are very similar, except for the "Sharp" parameter, which will be explained later.

| | Loud | Pleasant | Sharp | P4 | P5 | P6 | P7 |
|---|---|---|---|---|---|---|---|
| $R^2$ | 0.95 | 0.84 | 0.54 | 0.91 | 0.92 | 0.95 | 0.79 |

Table 1: Correlations between the probabilities obtained in the two multi-dimensional tests for each parameter

#### b. Comparison of the quality of results

The quality of the results can be evaluated in various ways. One can simply consider the repetition error, i.e. the difference of judgement between the first training pair and the same pair which had to be evaluated again later. The means of these differences (in absolute values) are shown in Table 2.

| | Loud | Pleasant | Sharp | P4 | P5 | P6 | P7 |
|---|---|---|---|---|---|---|---|
| Test 1 | 0.6 | 0.7 | 0.9 | 1.3 | 0.8 | 0.4 | 1.1 |
| Test 2 | 0.45 | 1.1 | 1.7 | 0.8 | 0.4 | 0.9 | 1.1 |

Table 2 : Means of the absolute values of differences between the training pair and its repetition

These differences are also very similar from one test to another and relatively slight (as a reminder, the marks go from -3 to +3). The only major deviation again concerns the "Sharp" characteristic. In fact, examining the results more closely, it can be seen that this deviation is strongly affected by two listeners in test 2, who had difficulties in evaluating this parameter. If they are removed from the sample, the repetition difference would fall from 1.7 to 1.1, which is a value similar to those for the other parameters. When computing the proportions for the "Sharp" characteristic without these two listeners, the correlation ($R_2$) with the proportions obtained in test 1 increases from 0.54 to 0.70.

Another way of evaluating the quality of the findings is to calculate a mean rate of inconsistencies, as follows. Consider a triplet of noises for which there are 3 marks, $n_{ij}$, $n_{jk}$ and $n_{ik}$. Let $d_{ijk} = n_{ij} + n_{jk} - n_{ik}$; if the listener has made a perfectly orderly judgement, which seems realistic at least for the parameters other than "Pleasant", we shall have $d_{ijk} = 0$. As the scale is limited between -3 and +3, it was decided to clip $n_{ij}+n_{jk}$ before substracting $n_{ik}$; therefore the expression is :

$$d_{ijk} = \max(-3; \min(3; n_{ij}+n_{jk}) )-n_{ik} \qquad (1)$$

And the overall consistency criterion is

$$C = \sqrt{\frac{1}{P}.\sum_{i,j,k} d_{ijk}^2} \qquad (2)$$

where P is the number of ordered triads (i,j,k).

This indication of the precision of a listener's judgement can be averaged over the whole jury for each test, to give the mean precision levels shown in Table 3.

| | Loud | Pleasant | Sharp | P4 | P5 | P6 | P7 |
|---|---|---|---|---|---|---|---|
| Test 1 | 1.48 | 1.75 | 1.48 | 1.96 | 1.93 | 1.72 | 2.11 |
| Test 2 | 1.35 | 1.44 | 1.92 | 1.45 | 1.92 | 1.33 | 1.34 |

Table 3: Mean precision of judgements

These precision levels are always slightly lower in the case of test 2, except in the case of the "Sharp" characteristic. Once again, if the two hearers who had problems with this parameter are removed from the panel, the mean precision becomes 1.41, a value similar to that for test 1.

### c. Comparison of length and difficulty of tests

The average durations of the tests and the average numbers of listening trials are shown in Table 4.

| | Test 1 | Test 2 |
|---|---|---|
| Length of test | 37 min. | 27 min. |
| Number of hearings | 91 | 104.5 |

Table 4 : Test length and average number of hearings

The second procedure allows shorter tests, in spite of the larger number of trials (note that the difference between these numbers of trials is not significant). The replies are therefore given more quickly, which is normal because the listener is concentrating on the parameter to be evaluated.

It should be noted that, as we have 11 pairs (10 plus the training pair) and 7 parameters, the minimum number of hearings is 11 for test 1 (assuming that a single hearing enables evaluation of the differences according to the whole set of parameters), and 77 for test 2. The average number of trials for test 1 shows that the subjects do not hesitate to repeat the pairs so as to perform the requested task satisfactorily.

The questionnaires did not show that one of the procedures seemed longer or more difficult than the other to the subjects.

### d. Conclusion

It can be said that the two methods give equivalent results. However, the second one seems to give more reliable results while enabling a shorter test time (although with a larger number of listening trials). It is therefore preferable, for a multi-dimensional test, to compare all the pairs for each parameter, rather than explore all the parameters for each pair individually.

## 4.B. Analysis of overall findings

### a. Differences between listeners

The answers of the whole panel of listeners (21) were examined in order to determine if there were inter-subjects differences. It was verified that, for any parameter, it was not possible to group these listeners in a set of uniform classes (by hierarchic classification, determined by means of the StatLab software). That means that all subjects made their evaluations in the more or less same way.

### b. Construction of absolute scores

Then, for each parameter, the merit values of the 5 noises was computed by the Bradley-Terry-Luce technique[13]. These merit values give a one-dimensional hierarchy between sounds. They can easily be computed as follows: if $P_{ij}$ are the classification proportions, the merit value of noise i is

$$v_1 = \sum_{j \neq l} \mathrm{Ln}\left(\frac{\mathrm{P}_{ij}}{\mathrm{P}_{ji}}\right) \qquad (3)$$

The underlying assumption here is that the results of each parameter are indeed one-dimensional. A way of validating this assumption[14] consists of computing back proportion estimators $\tilde{\mathrm{P}}_{ij}$ from the merit values, where

$$\tilde{\mathrm{P}}_{ij} = \frac{1}{2}.\tanh\left(\frac{v_i - v_j}{2}\right) \qquad (4)$$

and to compare these estimators with the real proportions. For each parameter in the present study, the coefficient of correlation between the experimental proportions and the estimators calculated according to the BTL values was always greater than 0.97, which shows a good match with the model.

It can therefore be considered that the judgements of the listeners according to each parameter indeed took place in a one-dimensional mode. Moreover, since the scatter between subjects is low, this means that the semantic ambiguities were reduced to the point of introducing no extraneous variability into the results.

### c. Comparison with findings based on analysis of verbal comments

For each parameter, the above mentioned BTL merit values were compared to the ($F_i$) values obtained from the free verbalisations (as explained in section 2.C). The correlations between these two sets of scores are shown in Table 5. All of these correlations are highly significant.

|  | Loud | Pleasant | Sharp | P4 | P5 | P6 | P7 |
|---|---|---|---|---|---|---|---|
| $R^2$ | 0.95 | 0.81 | 0.82 | 0.90 | 0.84 | 0.94 | 1.0 |

Table 5 : Correlations between BTL and verbalisation scores

The similarity between the two sets of scores can be seen in Figure 3 ; in this figure, the ($F_i$) values were normalised with respect to their mean, in order to be zero-centred as the merit values.

Subjects in the second test series therefore correctly interpreted the meaning of the parameters highlighted during the free verbal comments.

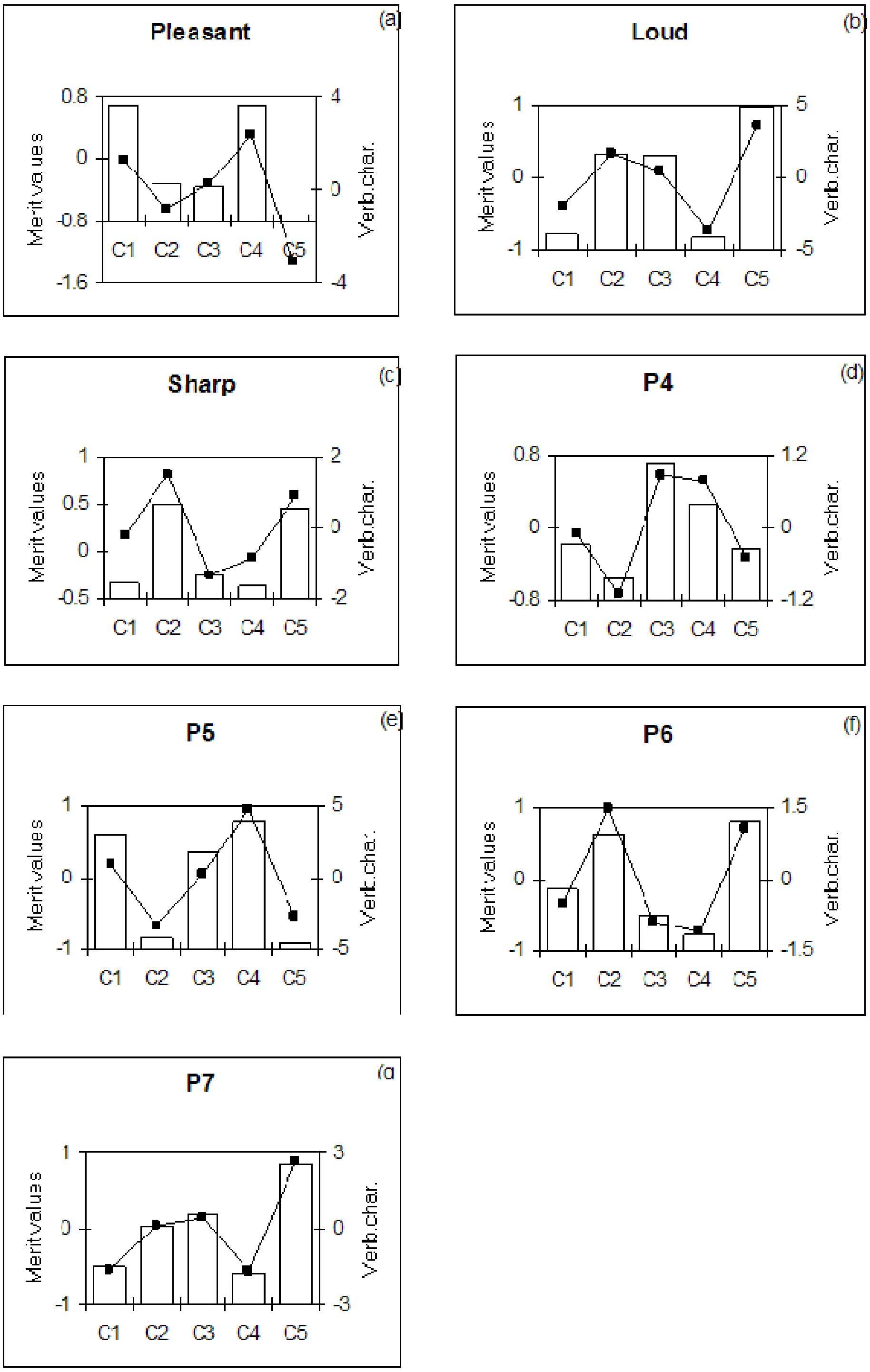


**Figure 3** : Comparison of the characterisations obtained from verbal comments (line) and merit values computed from the results of multi-dimensional test (bars).

## d. Conclusions

The very short learning process, using only verbal explanations (and no sound presentations) was enough to ensure a good understanding of the parameters by the listeners; the evaluations were made on a single scale for each parameter. Therefore, this set of parameters can be used in a multi-dimensional listening test and provide reliable information about the subjective evaluation of automobile diesel noises.

## 5. GENERAL CONCLUSION

It is therefore possible to build, by analysis of free verbal comments, a base of sound parameters which can be used in a multi-dimensional listening test. With a very short learning process, which merely presents to the subjects excerpts of verbal comments to highlight the meaning of the various parameters, semantic ambiguities are avoided and the results are reliable. Of course, the data base thus obtained can be used only in a sound context similar to that used for the analysis of the verbal comments. In our case, since it was built for automobile diesel engine idling noises, it is not recommended that it be used to evaluate high-speed driving noises, for example. Nevertheless, it could be advisable to build a set of data bases for a number of sound situations commonly encountered in road vehicles, which would allow multi-dimensional tests to be carried out in routine fashion.

Moreover, in a multi-dimensional test, one can recommend the procedure by which each parameter is considered in succession, and for which subjects are asked to compare all the pairs of noises. This procedure gives the same results as the commonly used procedure (by which noises are compared for each pair according to all the descriptors before considering the following pair), while offering improved precision and allowing shorter tests.

## 6. AKNOWLEDGEMENTS

The authors are grateful to Dr. Lawrence Finegold for assistance with English expression in the manuscripts.

## 7. REFERENCES

1 McAdams S., Winsberg S., Donnadieu S., De Soete G., Krimphoff J. “Perceptual scaling of synthesized musical timbres : common dimensions, specificities and latent subject classes” Psychological Research **58**, 177-192 (1995).

2 Grey J.M. “Multidimensional perceptual scaling of musical timbres” J. Acoust. Soc. Am. **61**, 1270-1277 (1997).

3 Takao H., Hashimoto T. “Subjective evaluation of running car interior noise - Part I. Selection of pairs of adjectives for the semantic differentials” Jidosha Gijitsukai Rombunshu **42**, 73-78 (1989).

4 Bisping R. “Car interior sound : experimental analysis by synthesis” Acustica - acta acustica **83**, 813-818 (1997).

5 Takanami K., Iwahara M., Saito H., Sakata M. “Improving interior noise produced during acceleration” SAE paper n° 911078 (1991).

6 Blommer M.A., Amman S.A., Otto N.C. “The effect of powertrain sound on perceived vehicle performance” SAE paper 971983 (1997).

7 Parizet E., Deumier S., Milland E. “Car road noise annoyance : significant timbre parameters and inter-individual variability” Proc. Forum Acusticum 1996, Acustica - acta acustica **82**, S216  (1996).

8 Nosulenko V.N, Samoylenko E.S. “Approche systémique de l’analyse des verbalisations dans le cadre de l’étude des processus perceptifs et cognitifs Social Sciences Information **36**(2), 223-261 (1997)  .

9 Nosulenko V.N., Parizet E., Samoylenko E.S.“Application de la méthode d’analyse des verbalisations pour caractériser les bruits de véhicules”, to be published in Social Sciences Information (1998).

10 Samoylenko E., McAdams S., Nosulenko V. “Systematic analysis of verbalizations produced in comparing musical timbres” Int. Journ. of Psychology **31**(6), 255-278 (1996).

11 Russel M.F. “An objective approach to vehicle internal noise assessments” Proc. Autotech 93 C462/205 (1993).

12 Schiffbänker H., Brandl F.K., Thien G.E. “Development and application of an evaluation technique to assess the subjective character of engine noise” Proc. SAE Noise and Vibration Conference 911081 (1991).

13 McGuire D.P., Davison M.I. “Testing group differences in paired comparisons data” Psychological bulletin, **110**(1), 171-182 (1991).

14 Amman S., Greenberg J. “Jury evaluation and quantification of automobile strut noise”, Proc. SQS-98, 147-152 (1998).