Comparison of some listening test methods : a case study

Etienne Parizet, Nacer Hamzaoui, Guillaume Sabatie

Acta Acustica united with Acustica 91 (2005), 356-364

Comparison of some listening test methods : a case study

E. Parizet, N. Hamzaoui, G. Sabatié
Laboratoire Vibrations Acoustique, Insa Lyon
25 bis avenue Jean Capelle, 69621 Villeurbanne Cédex, France
email : parizet@lva.insa-lyon.fr

Abstract : The goal of this study was to compare various listening test methods in the particular case of nine in-car ventilation noises. Six listening tests were conducted :
- absolute evaluation of noise pleasantness;
- evaluation of pleasantness, during which the subject could hear all noises as often as necessary;
- paired comparisons (forced choice procedure);
- paired comparisons (five levels scale);
- paired comparisons (continuous scale);
- similarity ratings, analysed with a multidimensional scaling method (Indscal).

These six tests were realised by 64 subjects.
Different items were examined for each test : its perceived and real duration, its estimated difficulty, the accuracy of merit scores attributed to noises, the perceptual spaces which could be built and the accuracy of a pleasantness indicator computed from the results.
It appeared that :
- the second procedure can propose a good compromise between the accuracy of the results and the time needed for subjects to realise the test. Thus, it can be recommended for many industrial purposes;
- however, the not-forced choice paired comparisons enable a greater discrimination between stimuli;
- perceptual spaces built from the paired comparison tests and the similarity rating one were similar, indicating a great stability of sound features used by listeners, whatever their task.

1 Introduction

The research of a better sound quality is from now on well accepted by manufacturers of industrial objects ("sound quality" meaning that the image of the product transmitted by its sound is in accordance with the global image that the manufacturer wants to give to this product). Listening tests constitute an essential tool to reach this objective, considering the complexity of timbre elements interfering in both the appreciation of this sound quality and the variability of listeners' expectations.

Generally, a listening test can have several objectives :
- Firstly, to assess or compare the pleasantness of different sounds. These sounds can come from an object and a selection of its competitors, or different possible modifications of a prototype ;
- Then, to identify the timbre aspects used by listeners to evaluate the pleasantness. This knowledge can enable the designer to identify possible ways of improvements of the product if he knows how to link a physical modification to a timbre change ;
- Lastly, a criterion of the pleasantness of this type of noise can be defined, if sound indicators correctly representing these timbre aspects can be identified (indicators implying measurable data from recorded signals). Target values of this criterion can at last be used as a specification for the sound of a new product.

Numerous experimental procedures have been suggested to reach these objectives. Listeners can only evaluate sound pleasantness, using a magnitude estimation method, or a graduated scale, or by various comparison methods. This evaluation can appear along with a quantification of some sound descriptors (method of semantic differentials [1]), the comparison between the answers for the different descriptors allowing to explain the pleasantness answers. Sometimes, a first test consists in quantifying similarities between sounds ; the multidimensional analysis of the answers provides a perceptive space, used to explain the pleasantness, evaluated through a second experiment [2,3]. It is also possible to analyze free verbalizations of listeners [4] or to ask to these ones to categorize the stimuli [5] in a more psychological than psychophysical field.
This list is not at all an exhaustive one : its function is to illustrate the great variety of the suggested techniques to reach the objectives of a perceptive study.

While a great number of psychophysical methods have been compared for some fundamental psychoacoustic matters (threshold level, just noticeable difference, loudness of pure tones etc.), that was not the case for the evaluation of pleasantness of real sounds or for the determination of the perceptual space in which such sounds are heard. Kendall and Carterette [6], in the case of wind instrument sounds, have shown the benefits of the verbal attribute magnitude estimation methods (in which unipolar answering scales are anchored by the same attribute (e.g. loud / not loud) as opposed to the classical semantic differential procedure using bipolar scale (with opposite terms, e.g. loud / soft); the VAME methods gave a better differentiation of sounds.
For industrial noises, Rossi *et al.* [7] have evaluated the pleasantness of door closing sounds and sounds recorded in a driving car through paired comparison tests and a magnitude estimation one. They concluded that, even though the general tendencies shown by the two methods were the same, some discrepancies remained in the results.
Such a conclusion was not confirmed by Zeitler and Hellbrück [8] who used bipolar and unipolar scales, and magnitude estimations with or without reference to evaluate the pleasantness of sounds originating from various sources (musical instruments, home appliances, car engines or environmental sounds). Results given by these methods were very similar; but it should be noted that the very wide range of sounds could make such an agreement easier.
In the precise case of a Diesel engine idle noise, Parizet and Nosulenko [9] used sound descriptors, derived from the analysis of a verbalization test, to define different scales in a paired comparison test. The various sound descriptions, as given by the analysis of verbalizations and their measurement on the scales, were also quite similar.

Considering that there is a lack of comparison between pleasantness evaluation methods, the goal of this study was to make such a comparison, in the precise case of the car ventilation system noise and for some of the available methods only. More precisely, two questions were asked :

- can some listening test procedures be compared as regard to their accuracy in the evaluation of the pleasantness of sounds, the difficulty of the listeners task and the duration of the test session ?
- is it possible to identify the perceptual space of sounds through pleasantness evaluation or comparison only, without conducting a similarity rating test ? If this is true, the overall duration of the experiment can be considerably reduced.

2. Experiments

2.1. Simuli

Nine auditory signals were used. They were recorded through a dummy head placed on the driver's seat of four luxury cars, while the engine was off and the ventilation system of the cabin was running in different settings (heating or air conditioning, various fan rotational speeds, etc). In the whole set of recordings, the nine selected signals have similar loudness levels : the maximum difference between them is less than 1 dB(A), or 1 Phone when using the ISO 532B loudness calculation. Actually, significant loudness differences could have lead listeners to evaluate or compare sounds on the basis of loudness alone, regardless of the test procedure applied. That would have prevented us from comparing these procedures; therefore, it has been decided to use sounds with nearly equal loudness values.
Each of the nine selected signals had a duration of 10 seconds, including an initial and a final 200 ms fading. They were presented to listeners through headphones (Sennheiser HD600), in a listening room isolated from exterior noise, at an average level of 74 dB(A).

2.2. Test procedures

Signals were evaluated through six different test procedures T1 – T6. Five of them were focused on the evaluation of the pleasantness of sounds :

- **T1** : sounds were presented to the listener one by one. First of all, the listener had to listen to all sounds at least once, in order to appreciate the context of the test. Then the evaluation began: after listening to each sound (as many times as necessary), the listener was asked to evaluate its pleasantness on a continuous scale, going from "very unpleasant" to "very pleasant". The scale was presented on the computer's screen and the answer was given by moving a cursor with the mouse. Sounds were presented to the listener in a random order;
- **T2** : sounds had to be evaluated using the same scale. The difference is that the nine scales were presented on the screen. Beside each scale was a button allowing the listener to hear the corresponding sound. In that way, the listener could compare the different stimuli before giving an answer for each of them. This was a mixed procedure, between evaluation and comparison, which had been introduced independently by Bodden [10] and Maunder [11] in the field of noise evaluation;
- **T3** : this test was a forced choice paired comparison; after listening to a pair of sounds (separated by a 500 ms silence), the subject had to choose the most pleasant one, so that only two answers were proposed to him. The set of pairs (their number being 36) was ordered according to Ross series [12], after a preliminary random arrangement of the nine sounds. The listener was still presented the whole set of sounds before starting the test and could hear each pair as often as he wanted to;
- **T4** was also a paired comparison; the only difference with the previous test was that the listener had the choice between five answers : "sound A is much more pleasant", "A is more pleasant", "A and B are equally pleasant" and so on;
- **T5** was another paired comparison test, the answer being given on a continuous scale. The scale was divided into four equally wide intervals with the same categories as used in T4 denoting the interval limits. The rest of the procedure was the same as for the third and fourth tests.

The sixth procedure **T6** was dedicated to the rating of the similarity of sounds within a pair. The listener had to give his answer on a continuous scale, the extremities of which are labeled "sounds are very similar" and "sounds are very different". The ordering of the sound pairs was the same as the one used for the paired comparison tests.
Each of the different test procedures was conducted from a computer, which recorded the answer (as a number between 0 and 1), the number of listenings of each sound (or pair of sounds) and the duration of the test.

Figure 1 shows the different answering scales of these six listening tests.

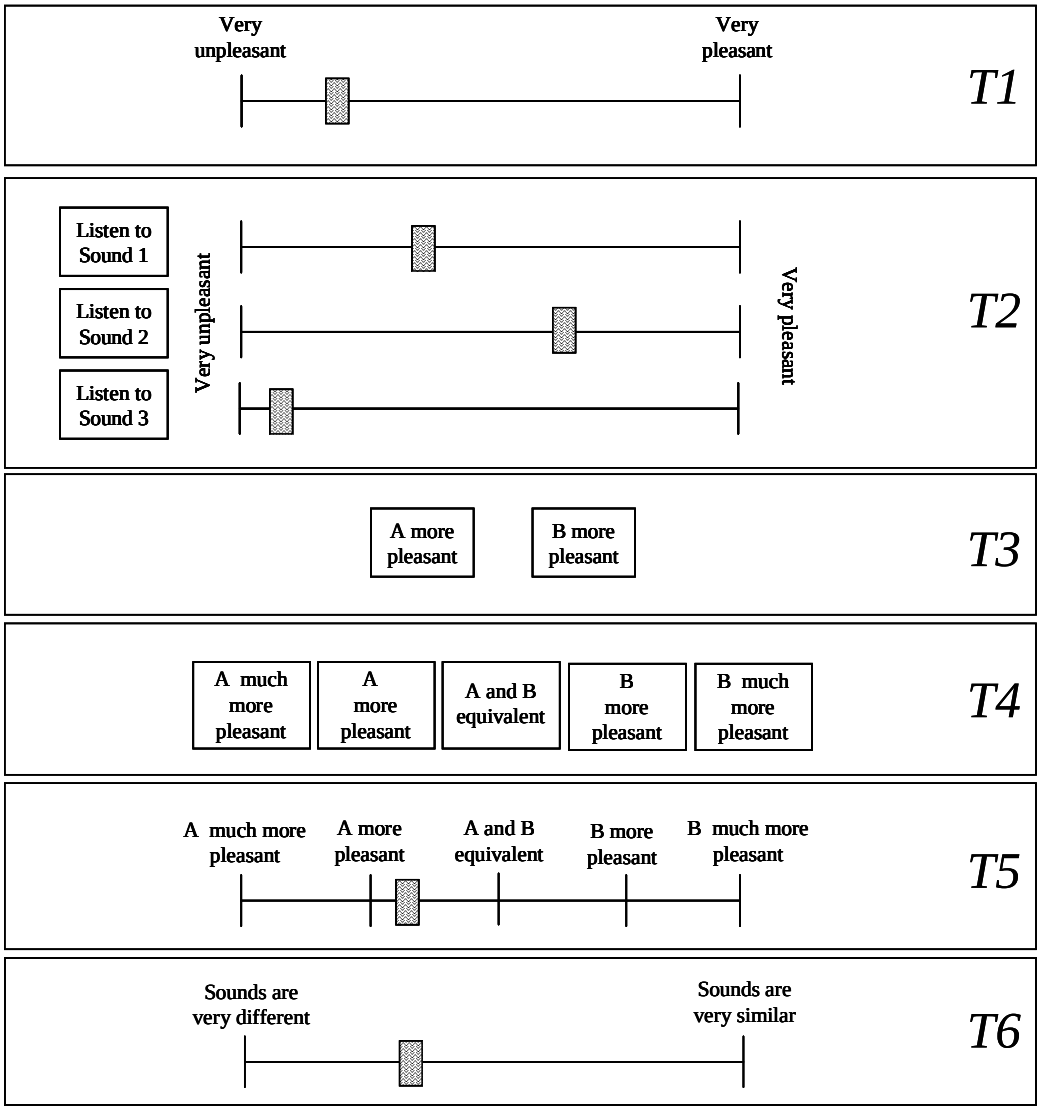


Figure 1 : answering scales of the six listening tests (for T2, only three out of nine scales are shown).

2.3. Subjects

Optimally, each listener would have achieved the six listening tests in a completely balanced order. That was not possible, because the number of listeners would have been far too high (it would amount to 6! = 720). So it was decided to simplify the strategy in the following way :

- each subject participated in two test sessions. Each session consisted of three tests and both sessions were separated by one week;
- the two evaluation tests T1 and T2 belonged to two different sessions. As the tasks required by these two procedures were the same, the risk would have been a great similarity between the answers;
- for the same reason, the paired comparisons T4 and T5 (for which the answers were given on a five levels scale or on a continuous one) did not belong to the same session;
- lastly, T1 and T2 were passed at the beginning or at the end of a session, in order to appreciate the influence of the noise habituation on their evaluation.

In such a way, the number of possible combinations was reduced to 64, which is affordable. 64 subjects had therefore participated to the experiment, each of them being submitted to the six listening tests. Most of these subjects were students, their age varying between 22 and 46 (average 24); 46 of them were male. They were paid 15 Euros for their participation.
After each test, the listener was asked to evaluate its length and difficulty. The answers were given on continuous scales graduated in five levels, going respectively from "very short" to "very long" and from "very difficult" to "very easy" respectively. Then he had a rest for some minutes before going on for the next listening test. After having achieved the first session, he was asked to come back one week later for the second session.

3 Results

3.1. Test duration

In figure 2 the averaged estimated duration of each test is shown. T1 and T2 (evaluation of sounds) were estimated as "rather short" ones by listeners, whereas the paired comparison ones seemed to be "rather long". An analysis of variance showed that the estimated duration of T1 and T2 were different at the confidence level of $p<0.01$, which was not the case between the paired comparison tests.
The averaged real duration of the test is also presented on the same figure (for each subject, the duration was recorded by the computer running the test). The narrow relation between physical and subjective duration is obvious : the upper limit for a test not to be evaluated as a long one is about 15 minutes.

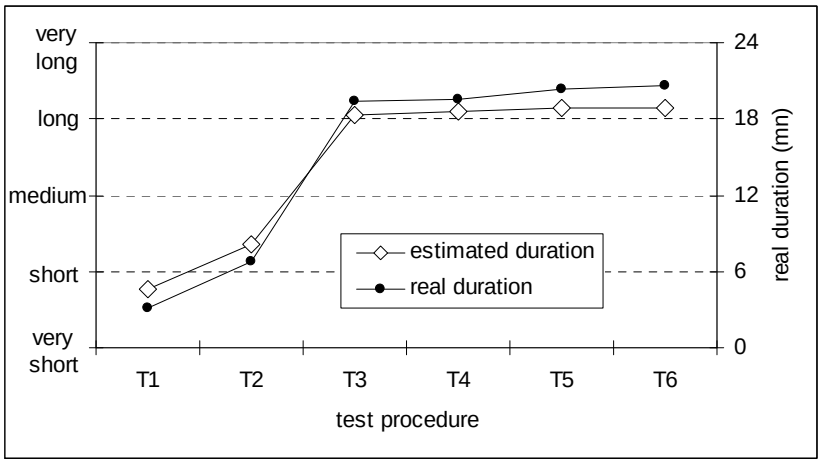


Figure 2 : estimated and physical duration of the six listening tests

3.2. Test difficulty

Figure 3 shows the average estimated difficulty of each test. These results are very close to one another ; but an ANOVA revealed that T6 (similarity ratings) was more difficult than all the other ones (at $p<0.05$ for T5 and $p<0.01$ for the first four tests). The rating of similarity is a more difficult task than the comparison or the evaluation of pleasantness.

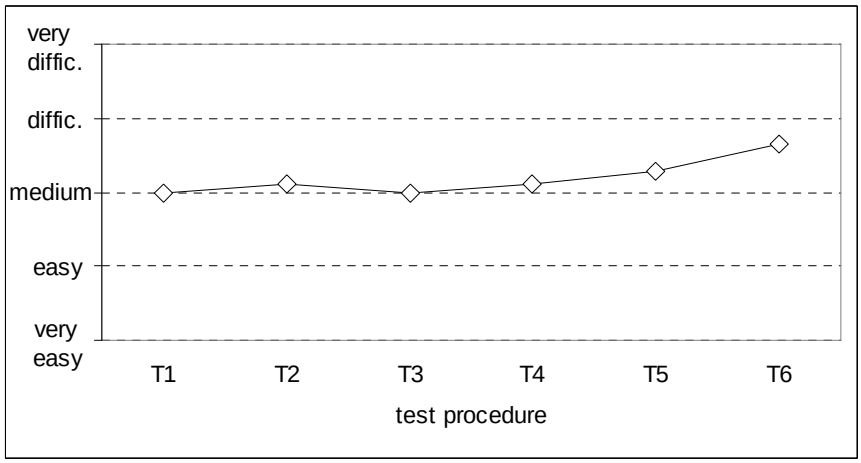


Figure 3 : estimated difficulty of the six listening tests

For a paired comparison test, another way of estimating its difficulty consists in computing the number of circular triads. For a forced choice procedure as in T3, this number can be computed from [12]

$$c = \frac{t}{24}(t^2 - 1) - \frac{T}{2} \qquad (1)$$

where

t is the number of sounds (here 9);

$T = \sum_{i=1}^{9} (a_i - \bar{a})$, where a_i is the merit score of sound i, and $\bar{a}$ is the average of the merit scores (respecting the relation $\bar{a} = \frac{t(t-1)}{2}$).

For T4 and T5, the number of circular triads was computed by examining all the possible sounds triads, as described in [13].

The averaged rates of circular triads were of 7.9 % for T3, 5 % for T4 and 9.7 % for T5. Therefore, the 5-level scale procedure allowed to reduce the number of errors in triads; a one-way repeated measures ANOVA revealed that the difference between T4 and the other two procedures was significant ($t_{T4,T3}$ = -3.6 and $t_{T4,T5}$ = -5.3 , $p<0.01$). The fact that T4 gave less circular triads than the two other procedures cannot be explained : in [13], it is argued that a forced-choice test can give more circular triads because the listener may perceive the two stimuli as equally pleasant but has to select one of them. But that explanation does not hold for T5.

3.3. Merit scores

In the continuation of the study, the merit scores of the noises were examined after being computed from each of the five first tests. In the case of the first two tests, these scores were

directly computed from the individual answers, by averaging these answers. For the paired comparison tests, they were linearly computed as :

$$S_i = \sum_{j \neq i} P_{ij} \qquad (2)$$

where P_{ij} is the preference probability of noise i versus noise j. As nine sounds were compared, merit scores computed from equation (2) ranged between 0 and 8; therefore, the scores computed from the first two tests were multiplied by 8, to allow the comparison between them.

Merit scores can also be computed through other techniques, the most widely used being Thurstone's law of categorical judgements (case V) or BTL model. In that study, using these two ones gave results which were very similar to those obtained from equation (2) (the correlation coefficient between merit scores achieved from these different techniques was greater than 0.99 for each listening test). This happens in many cases, as soon as most of preference probabilities are not close to 0 nor 1. For that reason, and also because later on in the study individual merit scores had to be computed, the linear computation of merit scores was used.

In figure 5 merit scores computed for each test are presented: they were very close to one another, which indicated a good agreement of the answers.

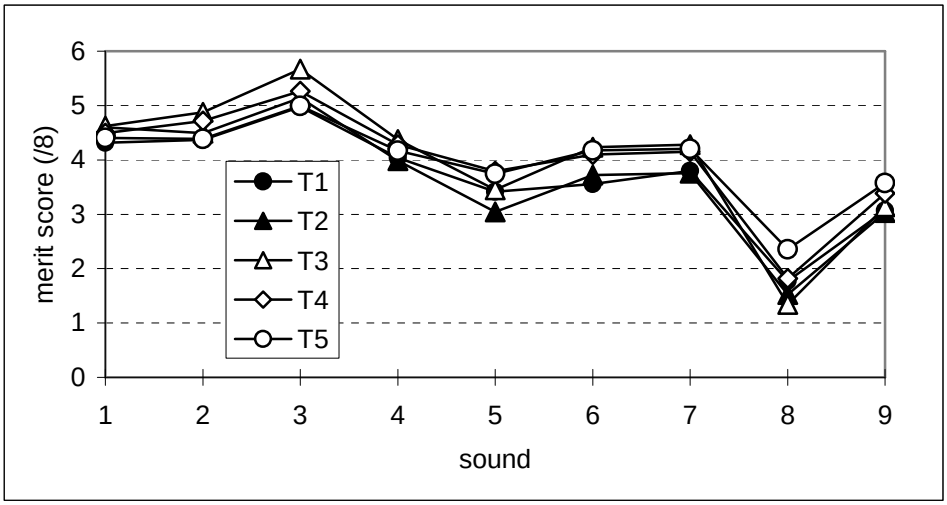


Figure 5 : merit scores computed from the results of the tests T1 to T5 (closed symbols : evaluation , open symbols : comparison)

The range of scores was greater for T3 than for the other paired comparison tests : this is understandable because T3 only allowed two answers. A clearly preferred sound was attributed the same answer by all listeners, whereas some listeners might express their preference in a weaker way in T4 or T5, by choosing a non-extreme answer.

On the other hand, the inter-individual variability of merit scores is higher for T3, as can be seen in figure 6, in which are drawn the standard deviations of sound merit scores. Two groups of listening tests can be clearly seen : the forced choice paired comparison gave standard deviations which were similar to those obtained from the two evaluation listening tests.

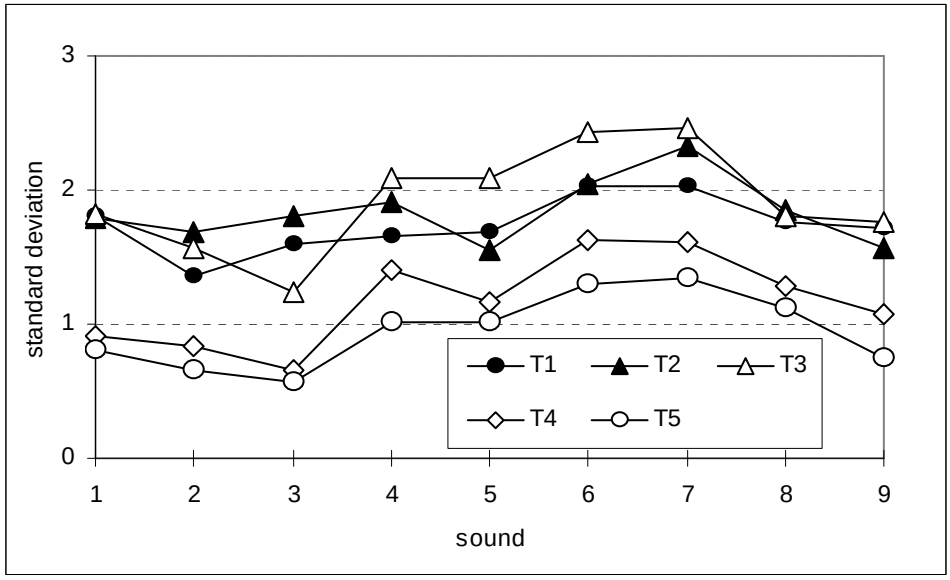


Figure 6 : standard deviation of merit scores for sounds

As a way of evaluating the discrimination power of each test, the pairs of sounds the scores of which were significantly different were counted using a one-way repeated measures ANOVA, at a confidence level of 5%. The maximum number of such pairs is 36 for 9 sounds : in this study, these numbers varied from 18 to 22 (table A). The maximum number was provided by T5 (paired comparison on a continuous scale), which indicated that this procedure allowed to discriminate sounds in a slightly better way than the other ones.

T1	T2	T3	T4	T5
18	18	19	21	22

Table A : number of pairs in which the merit scores of sounds are significantly different ($p < 0.05$).

In the particular case of T1 and T2, as mentioned before, half of the jury achieved T1 at the beginning of a session and T2 at the end of the other session, while the reverse order was used for the other half of the jury. An analysis of variance revealed that, in the case of T1, the results could be different if the listener was submitted to this test at the beginning or at the end of a session. Differences were highly significant ($p<0.01$) for sounds 3 and 8, which were attributed the best and the worst merit scores (see figure 5). Therefore, the experience of listening to sounds (while accomplishing the paired comparison tests) could make listeners use a greater range of the scale, though these listeners had to listen to all sounds at least once before performing the evaluation task of T1.

That effect could not be found in the results of T2 : no significant difference could be found in the results of the two halves of the jury. The possible comparison between sounds, allowed by that procedure, reduced the influence of the experience of sounds.

3.4. clustering listeners

A more precise analysis of the merit scores consisted in looking if the jury could be separated in groups of listeners with similar evaluation. For the five first tests, that was realized using the K-means technique [14]; the input data of the algorithm were the individual merit scores of noises, either directly obtained in the case of the first two tests, or computed using equation

(2) for T3, T4 and T5. In all cases, it appeared that a correct clustering (i.e. for which the number of people in each class is not too small) consisted in separating the jury in two groups. The number of listeners belonging to each group was quite similar for each test (table B).

	T1	T2	T3	T4	T5
Group 1	49	41	35	40	45
Group 2	15	23	29	24	19

Table B : number of listeners in each group of the two-class clustering of the results of tests T1 to T5.

Moreover, 34 listeners out of 64 always belonged to the same group and 18 other ones belonged to the same group for four of the five tests. This shows once again the great stability of the results.

For the first five tests, it was possible to make a comparison between the different merit scores computed from each group of listeners (figure 7 and 8).

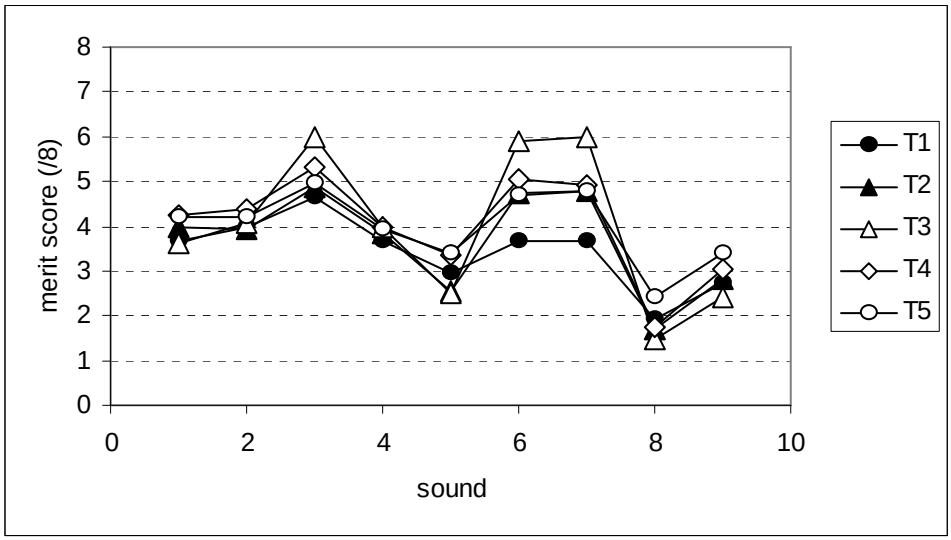


Figure 7 : merit scores of noises, computed from groups 1 of each test (T1 to T5)

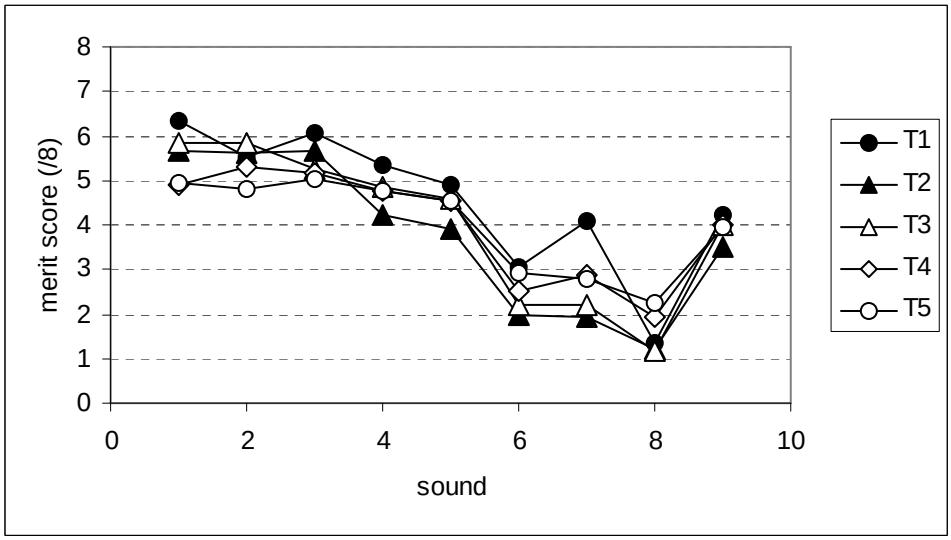


Figure 8 : merit scores of noises, computed from groups 2 of each test (T1 to T5)

Generally, differences between groups were related to sounds 1, 2, 5, 6 and 9. As it was the case for scores computed over the whole panel, results were very similar within the tests. The consequence of the forced choice test (T3) over the dynamic range of scores can be seen, especially in the case of the first group (figure 7). The evaluation test T1 gave results which were slightly different from the other ones for some sounds (sounds 6 and 7 for the first group, 7 for the second one). That reveals a lower accuracy of T1, when compared to the other procedures.
For each test, the statistical significance of differences of scores computed for the two groups of listeners was examined. Results are summarized in table C : they show that all paired comparison tests enabled a higher discrimination between the groups of listeners.

	T1	T2	T3	T4	T5
Sound 1	< 1%	< 5%	< 1%	< 1%	< 1%
Sound 2	< 5%	< 1%	< 1%	< 1%	< 1%
Sound 3	*ns*	*ns*	< 5%	*ns*	*ns*
Sound 4	*ns*	*ns*	*ns*	< 5%	< 1%
Sound 5	< 1%	*ns*	< 1%	< 1%	< 1%
Sound 6	*ns*	< 1%	< 1%	< 1%	< 1%
Sound 7	*ns*	< 1%	< 1%	< 1%	< 1%
Sound 8	*ns*	*ns*	*ns*	*ns*	*ns*
Sound 9	< 5%	*ns*	< 1%	< 1%	< 1%

Table C : level of significance for merit scores between groups of listeners

3.5. identification of the perceptual space and preference model

A classical way of determining the perceptual space consists in performing a multi-dimensional analysis of similarity ratings. Therefore, the results of T6 were analyzed with the INDSCAL procedure, as defined by Carroll and Chang [15]. The analysis was repeated for a number of dimensions of the perceptual space varying from 1 to 8; for each solution, Pearson's correlation coefficient between measured dissimilarities and re-constructed distances was computed. The four-dimensions solution gave a correlation coefficient of 0.88, which was considered to be satisfactory.
In figure 9 are shown the 1-2 and 3-4 planes of this perceptual space.

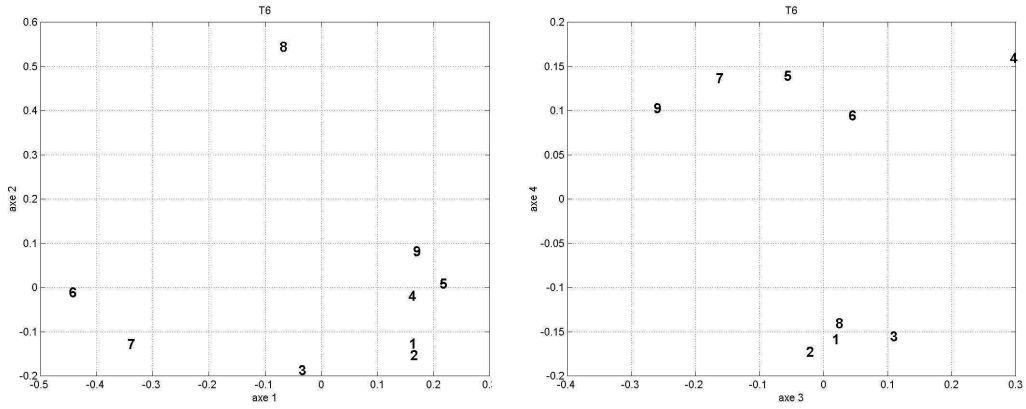


Figure 9 : Planes 1-2 and 3-4 of the perceptual space, obtained from an Indscal analysis of T6 similarity ratings

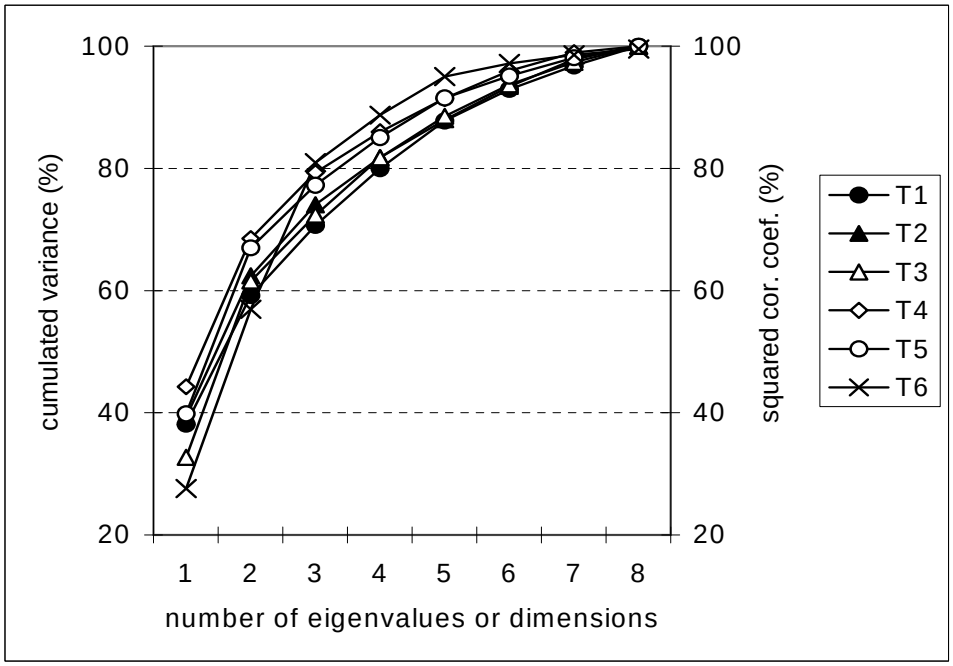
It can be seen of figure 9 that some of the sounds, which had been recorded in the same vehicle, were close to each other, at least on the first and second axis. This is the case for sounds 1 to 4 (recorded in the same car) and 6 and 7 (in a second one). In spite of the different

settings on the ventilation systems, there is a timbre similarity between them. On the other hand, sounds 8 and 9 , which had also been recorded in the same car, were far from each other (and sound 5 was recorded in a fourth car).
Listening to sounds could give information about features creating the first two axis :

- The first one was due to the spectrum balance of sounds. Sharp sounds are located on the right side of this axis, while the less sharp ones (6 and 7) are on the left side. The sharpness metric, computed from the ISO532-B specific loudness curves (by Mts Sound Quality® software), was significantly correlated with the co-ordinates of sounds upon the first axis (R = 0.99).
- The second one was essentially created by sound 8, that a sharp whistle (around 8 kHz) made different from other sounds. It should be noted that it was not possible to identify a sound metric correctly related to that axis : Tonality, Prominence Ratio and Tone-to-Noise Ratio did not exhibit a significant correlation with sounds co-ordinates over this second axis.

On the other hand, it was not possible to identify sound features related to the third and fourth axis.

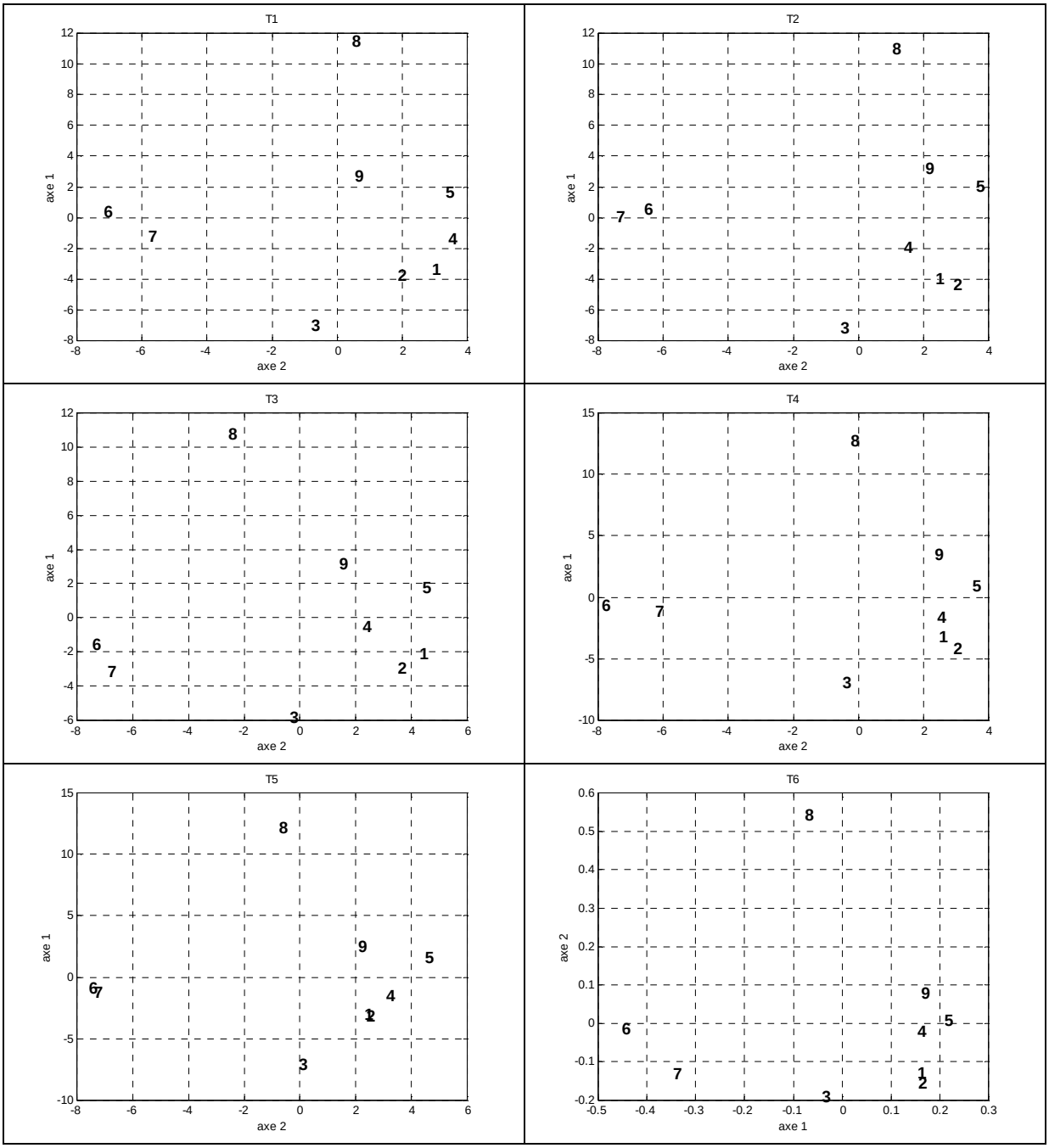
The hypothesis was made that it was possible to identify that perceptual space from results of the five first tests. Therefore, a Principal Component Analysis was conducted for each of these tests, in which data were the merit scores computed for the sounds from the results of each listener; variables of the analysis were listeners and individuals were sounds. In the case of the paired comparison tests, individual merit scores had been first computed using equation (2) for each listener. It should be noted that, for a paired comparison test, the number of degrees of freedom was only 8, because merit scores of the nine sounds respected the relation $\sum_i S_i = \frac{t(t-1)}{2}$, where $t = 9$.
In each case, the cumulated variance explained by the first four eigenvalues was greater than 80% (figure 10).



<u>Figure 10</u> : cumulated variance explained in the Principal Components Analysis of tests T1 to T5 and approximation criterion in T6

The comparison of the planes of the first two principal components of these analysis and the 1-2 plane computed from the Indscal analysis of T6 (figure 11) shows that these planes are very similar, though a swapping between first and second axis (it should be kept in mind that, for the five first tests, the first axis is the vertical one in figure 11). More precisely, the coefficient of correlation between the coordinates of sounds over each of the four principal components and those of sounds over the axis of the Indscal analysis was computed (table D). High values ($R^2 > 0.8$) indicate a good similarity within the 1-2 plane of each analysis; and even the third and fourth axis computed from T3 and T5 are close to those of the Indscal analysis.

This validates the hypothesis, as stated in [15], that similarity and preference judgements are based on the same acoustic parameters. But that result also points out that, as a similarity evaluation give no more information than a paired comparison test, it is not necessary to conduct it. As the length of these two tests are equivalent, that can considerably reduce the overall duration of the experiment.



<u>Figure 11</u> : perceptual spaces (axes 1 and 2) obtained from each of the six tests

	T6_1	T6_2	T6_3	T6_4
T1_1	-0.09	0.96	-0.22	0.13
T1_2	0.96	0.08	0.25	-0.24
T1_3	0.14	-0.22	-0.65	0.54
T1_4	0.05	0.02	-0.54	-0.71
T2_1	-0.14	0.93	-0.30	0.21
T2_2	0.96	0.15	0.10	-0.33
T2_3	-0.05	-0.11	-0.70	-0.42
T2_4	0.09	0.08	-0.20	-0.23
T3_1	0.13	0.97	-0.16	0.02
T3_2	0.98	-0.16	0.13	-0.25
T3_3	0.04	0.12	0.89	0.02
T3_4	0.02	-0.11	-0.01	0.87
T4_1	-0.06	0.97	-0.24	0.08
T4_2	0.99	0.04	0.10	-0.25
T4_3	0.01	0.10	0.75	0.48
T4_4	-0.05	0.17	0.26	-0.59
T5_1	-0.02	0.97	-0.22	0.07
T5_2	0.99	0.01	0.17	-0.20
T5_3	-0.11	0.15	0.90	0.21
T5_4	0.05	-0.14	-0.31	0.90

Table D : correlation coefficients between the coordinates of sounds on each principal axis (computed from each of the five first tests) and those on the axis of the Indscal analysis

The last point of the study was related to the relation between co-ordinates of sounds on the first two axis of the perceptive space and their merit scores. As two groups of listeners could be identified, it was tried to compute a model of merit scores averaged over each group from co-ordinates of sounds over the first two axis. First of all, the perceptual space provided by the Indscal analysis of T6 was used. Merit scores obtained from the results of T5 were selected, because it had appeared that test discriminated sounds in the greatest way. In that case, models were very accurate :

$$\begin{cases} S_{group1}^{T5} = 4 - 1.6\, X_1^{T6} - 3.3\, X_2^{T6} \\ S_{group2}^{T5} = 4 + 3\, X_1^{T6} - 3\, X_2^{T6} \end{cases} \qquad (3)$$

the correlation coefficients between measured and predicted merit scores being of 0.95 for the two groups of subjects. Figure 12 shows the high accuracy of models defined by equation (3). The constant values in equation (3) are nearly 4, which is the average of scores (as T5 was a paired comparison test, that average is $\frac{t(t-1)}{2}.\frac{1}{t} = 4$).

Equations (3) emphasize the difference between listeners in each group, because the coefficients of co-ordinates of sounds upon the first axis are of opposite signs. As that axis was explained by the frequency balance in sounds, it can be said that listeners from the first

group preferred sounds with a low frequency content, whereas subjects from the second group preferred rather sharp sounds. That different appreciation of frequency balance had already been identified in other cases, dealing with road noise inside car cabins [16] or noise in a high speed train [17]. On the other hand, the sharp whistle of sound 8 was not appreciated by all listeners (the coefficients of X_2^{T6} are negatives for the two groups and have the same order of magnitude).

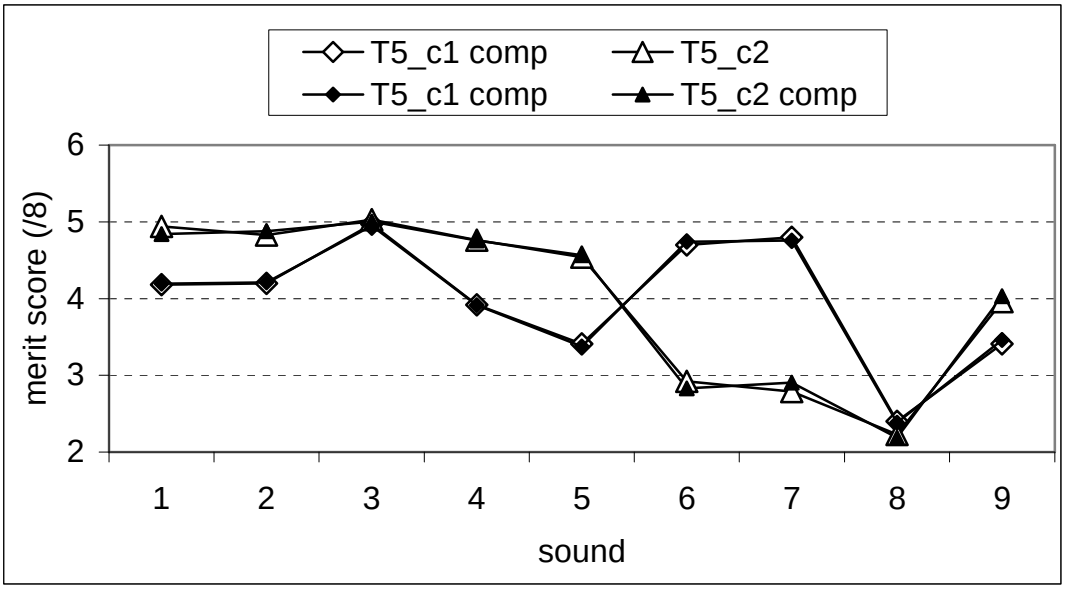


Figure 12 : merit scores computed from the results of T4 for the two groups of listeners and computed from equation (3). Open symbols : measured values; black symbols : values computed from eq. (3)

For evaluation or comparison tests (T1 to T5), as the same data were used to compute the merit scores and as inputs to the Principal Components Analysis, it was natural to find such a model. For example, in the case of T5, the two models were :

$$\begin{cases} S_{group1}^{T5} = 4 - 0.137\, X_1^{T5} - 0.084 X_2^{T5} \\ S_{group2}^{T5} = 4 - 0.140\, X_1^{T5} + 0.174\, X_2^{T5} \end{cases} \quad (4)$$

S^{T5} being the merit scores of sounds (computed for each group of listeners), X_1^{T5} and X_2^{T5} being sounds co-ordinates upon the first two principal axis of the PCA analysis of T4 results. The correlation coefficients between measured and predicted scores were more than 0.99 for the two groups. As the principal axis are swapped with those obtained from the Indscal analysis of T6, the difference between the two groups of listeners now appears on the coefficient of X_2^{T5}.

4. Conclusion

In the case of the car ventilation noises which was used for this study, it can be said that :

- the five tests dealing with the evaluation or the comparison of noise pleasantness had similar results, when these results were computed over the whole panel of listeners;
- the discrimination power was greater for paired comparison test than for the evaluation ones;
- within these paired comparisons, the procedures in which the equality answer is allowed gave less scattered results;
- the absolute evaluation test (T1) provided less accurate results, when these results were related to the separation of the panel into homogenous sub-groups. Also, previous noise presentations could modify subject's answers; this did not affect the results of the mixed evaluation test (T2);
- the perceptual spaces obtained from the various tests showed very strong similarities. The first two axis, are very stable and the continuous scale paired comparison test also gave very similar third and fourth axis.

In a practical way, the recommendations obtained from that study are that, if the goal of a listening test is only related to the evaluation of sounds pleasantness, the mixed evaluation test (T2) can offer a good compromise between the accuracy of the results and the time needed by each subject to achieve the test. Both that method and a paired comparison one, allow to understand the different appreciation of listeners. However, in order to maximize the discrimination power of the test, a not forced-choice paired comparison can be recommended. Moreover, such procedures can provide useful information about the perceptual space which can make a similarity evaluation not necessary.

It should be kept in mind that these conclusions are valid in the context of sounds being used for that study. For example, it should be reminded that loudness of sounds were more or less constant : if that had not been the case, differences between test procedures certainly would have been reduced, as loudness is a very important annoyance factor. It would certainly be useful to repeat that study with other types of noises, in order to check the validity of such conclusions.

5. Bibliography

[1] von Bismarck G. "Timbre of steady sounds : a factorial investigation of its verbal attributes", Acustica **30**(3) (1974), 119-182

[2] Susini P., McAdams S., A multidimensional technique for sound quality assessment, Acustica-Acta Acustica **85** (1999), 650-656

[3] Susini P., McAdams S., Winsberg S., Perry I., Vieillard S., Rodet X. "Characterizing the sound quality of air-conditioning noise", Applied Acoustics, 65 (8), (2004), 763-790

[4] Nosulenko V., Samoylenko E., Parizet E.,. "Evaluation and verbal comparison of noises produced by car engines" International Journal of Psychology **31** (1996), 3 – 4 (A)

[5] Guyot F. "Etude de la perception sonore en terme de reconnaissance et d'appréciation qualitative : une approche par la catégorisation", Ph.D. Université du Maine (Le Mans), 1996

[6] Kendall R., Carterette E. "Verbal attributes of simultaneous wind instrument timbres : I. VonBismarck's adjectives" Music Perception **10**(4) (1993), 445-468.

[7] Rossi G., Crenna F., Codda M. "Measurement of quantities depending upon perception by jury-test methods" Measurement **34** (2003), 57-66

[8] Zeitler A., Hellbrück J. "Psychophysical scaling of the pleasantness of everyday-noises", Proc. 8th Oldenburg symp. on psychological acoustics (2000), 233-240

[9] Parizet E., Nosulenko V. "Multi-dimensional listening test : selection of sound descriptors and design of the experiment", Noise Control Engineering Journal **47**(6) (1999), pp.1-6
[10] Bodden M. , Heinrichs R., Linow A., "Sound quality evaluation of interior vehicle noise using an efficient psychoacoustic method", Proc. Euronoise 98 (1998), 609-614
[11] Maunder R. "An interactive subjective assessment method for recorded sound" Proc. ImechE 1998, 345-354
[12] David H.A. "The method of paired comparison" Oxford University Press, New-York, 1988.
[13] Parizet E., "Paired comparison listening tests and circular error rates", Acustica-Acta Acustica 88 (2002), 594-598
[14] Anderberg M., "Cluster analysis for applications", Academic Press, New York, 1973.
[15] Carroll J., Chang J. "Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition", Psychometrica 35 (1970), 283-319
[16] Parizet E., Deumier S., Milland E., "Road noise subjective assessment in car"s, InterNoise, Liverpool 1996
[17] Parizet E., Hamzaoui N., Jacquemoud J. "Noise assessment in a high-speed train", Applied Acoustics **63** (2002), 1109-1124