Paired comparison listening tests and circular error rates

Etienne Parizet

Paired comparison listening tests and circular error rates

Etienne PARIZET
Laboratoire de Vibrations et d'Acoustique
Insa – Lyon
25 bis, avenue Jean Capelle
F-69621 Villeurbanne Cédex

tel : 33 (0)4 72 43 81 21
fax : 33 (0)4 72 43 82 17
email : parizet@lva.insa-lyon.fr

Summary

On the basis of noises recorded in trains, this article presents circular errors rates obtained by two possible paired comparison methods. The first one allows listeners to answer that two noises in a pair are equally annoying, while the second one forces them to select one of the sounds. It is shown that the latter can lead to very high number of circular errors when perceptual differences between noises are small. Therefore the former method is recommended, and the number of circular errors must be essentially considered as an indicator of the difficulty of the task.

1) Introduction

Listeners can make some mistakes on some triads of sounds in the tests of comparisons by pairs. These mistakes are composed of circular answers, which means that the listener expresses his preference to sound A than to sound B, sound B to sound C and also sound C to sound A. The origin of such mistakes is not well-known. According to Weber [1], they can be :

- A real inaccuracy from the listener who didn't pay enough attention to the test ;
- An alteration of assessment criteria during the test ;
- Sounds which were actually quite close to one another, making the task of the listener more difficult.

In the case of a forced choice test (in which the listener must choose one of the two noises), the number of circular triads can be computed for each listener by [2]

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

where

t is the number of stimuli ;

$T = \sum_{i=1}^{t} \left(a_i - \bar{a}\right)^2$, a_i being the score for the i noise for this listener and $\bar{a}$ being the average of scores ($\bar{a} = \frac{t-1}{2}$).

The coefficient of consistence is then defined by [3]

Ref : Acustica united with Acta Acustica 88 (2002), 594-598

$$\begin{cases} \zeta = 1 - \dfrac{24.c}{t(t^2-1)} & \text{if } t \text{ is uneven} \\ \zeta = 1 - \dfrac{24.c}{t(t^2-4)} & \text{if } t \text{ is even} \end{cases} \qquad (2)$$

This criterion allows to assess the reliability of the listener. In some previously published studies [1, 4, 5], listeners whose coefficient of consistence is lower than a fixed value are excluded from the analysis. The value can be set to 0.6 [4], 0.7 [1] or is not mentioned in the paper [5].

The purpose of this article is to show that a forced choice test can lead to low coefficients of consistence if stimuli are close enough to one another, because listeners make their choice at random. This is done on the practical example of noise in a high speed train.

2) Stimuli

Sounds were recorded through a dummy head (Bruel and Kjaer 4100) in a duplex high speed train (TGV). The dummy head was located either in the upper or in the lower room and the experiment was carried out while the train was running at different speeds on two types of tracks (the classical one or the high speedy one). Eight samples lasting 10 seconds each were taken into account.

3) First test (unequaled loudness)

3.1 Procedure

In the first experiment, sounds were presented to listeners at their real loudness levels; loudness differences among them were important ($\frac{N_{max}}{N_{min}} \approx 3$, where N_{max} is the loudness of the loudest sample and N_{min} the loudness of the quietest one, these values being computed according to the ISO 532B method implemented in the MTS *Sound Quality* software).
Sounds were submitted to a listener through an earphone (Sennheiser HD 600) in a quiet room. The method of comparisons by pairs was used with a scale consisting of 5 categories : the listener could answer that he found sound A much more annoying than sound B, more annoying, equally annoying, or, on the contrary, that B was more annoying than A or much more annoying.
Pairs were presented according to Ross's series [6] after a first aleatory permutation of the order of signals, which allows to avoid repetitions for the successive presentations of the pairs.
The overall test was achieved by a Matlab program on a PC computer which played sounds, presented the scale of answers to the listener, got and registered his answers.
48 listeners, who mainly were students, participated to this test, for which they were paid. None of them reported any hearing problem.

3.2 Results

The answers of each listener were converted into numbers between –1 and +1 (the possible values being –1, -0,5, 0, 0,5, 1). The accuracy in triads was considered in the following way for each listener :

If P_{12}, P_{13} and P_{23} were the answers when the presentations of pairs were made (1-2, 1-3 and 2-3), it can be considered that there was a circular mistake in two cases :

$$\begin{cases} P_{12} \geq A \text{ and } P_{23} \geq A \text{ and } P_{13} \leq -A \\ \qquad\qquad or \\ P_{12} \leq \text{-A and } P_{23} \leq -A \text{ and } P_{13} \geq A \end{cases} \tag{3}$$

A stands for a limit, the value of which establishes the accuracy allowing to study the results carefully : if A=0, no inverted preference can be accepted.

By considering the number of possible triads ($A_t^3 = \frac{t!}{3!}$, where t is the number of sounds), the rate of circular errors is defined as :

$$C = \frac{1}{A_t^3} \sum_{1 \leq i,j,k \leq t} \delta_{ijk} \tag{4}$$

where δ_{ijk} is equal to 0 or 1 according to (3).

It is worth noting that other indicators as

$$\begin{cases} \delta_{ijk} = Max\left[Min\left(P_{ij} + P_{jk}\,;1\right);-1\right] - P_{ik} \\ C = \frac{1}{A_t^3} \sum_{1 \leq i,j,k \leq t} \delta_{ijk} \end{cases} \tag{5}$$

can be used, the Max(Min) operator being here to take the limitation of scale of answers into account (the listener can then answer P_{ij}=1, P_{jk}=1 and P_{ik}=1 in a coherent way).

This way of computing the rate of circular mistakes is stricter than the previous one, because it means that the listener is expected to perfectly put in order the different sounds which are presented on the scale of preference.

Figure 1 stands for the values of the first indicator of coherence obtained from 48 listeners who were classified according to this indicator, A being fixed to 0.25. It means that a mistake is not taken into account if the listener thinks that two sounds out of the three studied ones are equally annoying. The values vary between 0 and 15%.

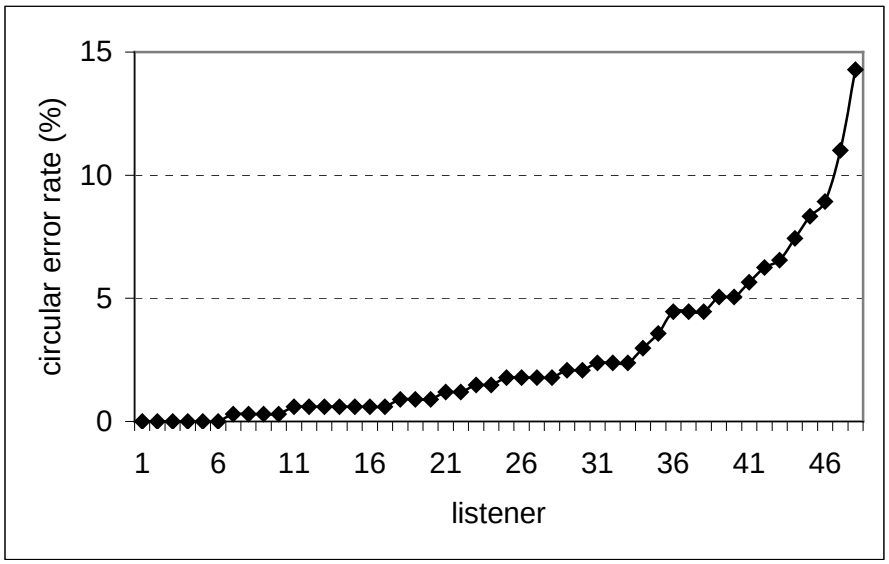

Figure 1 : Circular errors rates for original test procedure (original sounds)

Listeners can be divided into three groups :
- For the first one, C is lower than 1% (this group consists of 20 listeners);
- In the second one, ξ is lower than 2.5% (33 listeners);
- Lastly, the whole panel (48 listeners).

The merit scores were computed for each listener (by summing up the value of each row of his preference matrix) and then averaged over the three groups. These averaged merit scores are shown in figure 2 : the merit scores obtained over the three groups are very similar : therefore, it is not necessary to exclude some listeners from the panel.

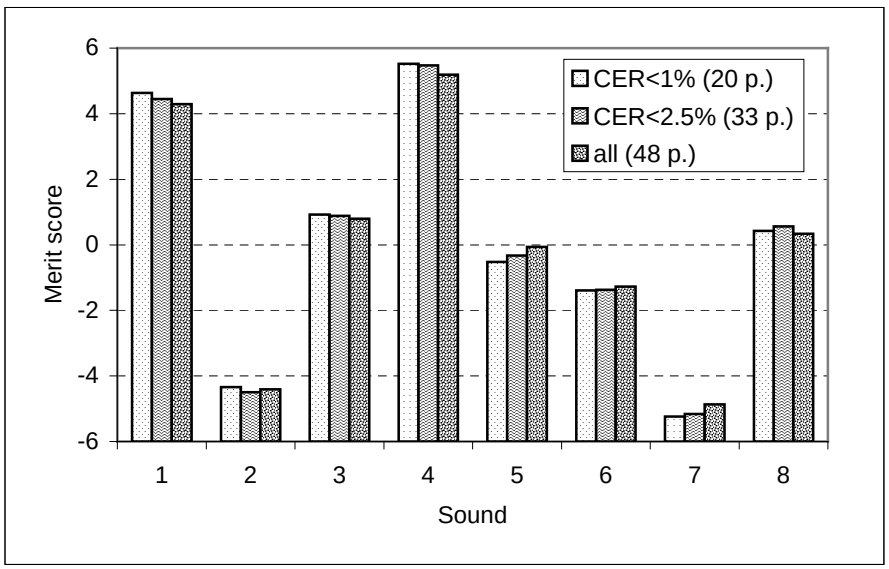

Figure 2 : Merit scores computed from differently reliable listeners results (original sounds)

3.3 Modification of the results

A ($\tilde{P}_{ij}^{(k)}$) chart can be determined from the chart of initial data ($P_{ij}^{(k)}$) (where i and j indicate sounds and k represents a listener), so as to show the results that could have been obtained through a forced choice test :

$$\begin{cases} P_{ij}^{(k)} > 0 \Rightarrow \tilde{P}_{ij}^{(k)} = 1 \\ P_{ij}^{(k)} < 0 \Rightarrow \tilde{P}_{ij}^{(k)} = 0 \\ P_{ij}^{(k)} = 0 \Rightarrow \tilde{P}_{ij}^{(k)} = \chi \end{cases} \quad (6)$$

where χ is a discrete random variable which can take the values 0 and 1 with an equal probability. It means that the listener would have chosen one of the two possible answers at random (A>B or B>A) if he had thought that both sounds would have been equivalent.

50 samples of the ($\tilde{P}_{ij}^{(k)}$) chart were thus determined, and for each sample, the coefficient of consistence of each listener was computed according to equations (1) and (2), leading to new individual random variable (ζ_k). The probability to get a coefficient of consistence higher than 0.7 was then computed for each listener (0.7 being the limit under which the listener can be taken off the panel). These probabilities are shown in figure 3.

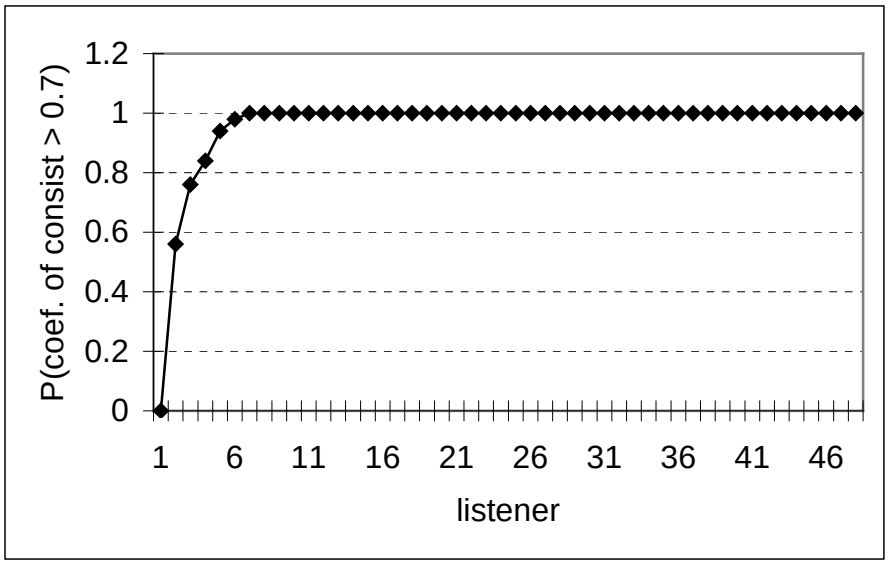

Figure 3 : Probability that ζ>0.7 for different listeners (original sounds)

The probability that ζ_k exceeds 0.7 is not one for only 6 people : that means that nearly all listeners would have been kept in the panel. Of course, there is a relation between this probability and the rate of circular errors computed from the real answers using equations (3) and (4), as it is shown in figure 4. The listener with the highest rate of circular error (C = 15 %) could not present a coefficient of consistence greater than 0.7 ($P(\zeta_k>0.7) = 0$).

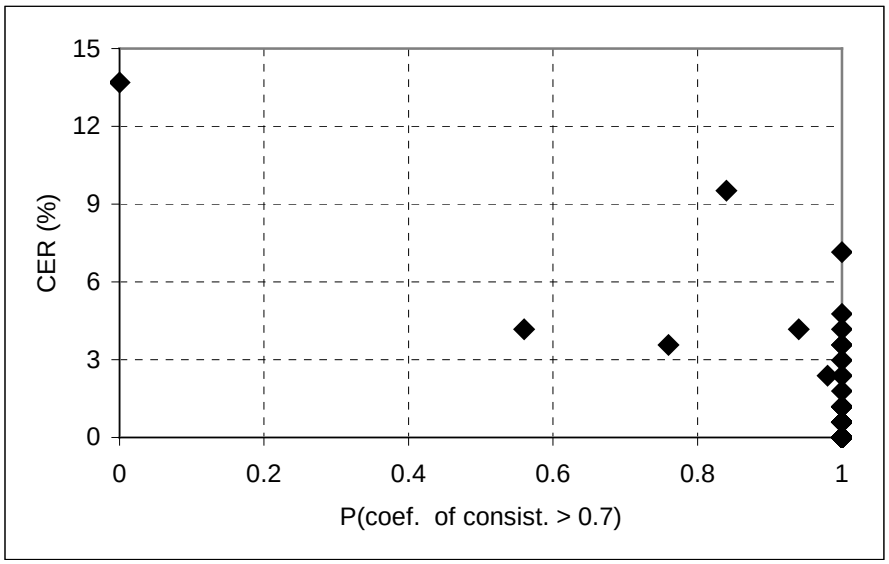

Figure 4 : Relation between CER (test with equality allowed) and probability that ζ>0.7 (forced choice test) for the original sounds

This confirms that most listeners succeeded in the rather easy task of comparing the original noises, because of loudness differences between these noises.

4) Second test (equalized loudness)

4.1 Procedure

The use of loudness determined the overall perceptive results obtained in the test described above. An equalization was made (by modifying the amplitude of signals so as to obtain an equivalent loudness, computed according to the ISO 532B norm, through the MTS *Sound Quality* software) in order to eliminate the influence of this parameter.
The same listening test procedure was then used, but this time each listener did it twice. 94 results are thus available (47 listeners).

4.2 Results

The individual rates of circular errors, computed according to (3), are much more important in the studied case than in the previous test (figure 5) ; they vary between 0 and 30 % and only a third of the listeners have a rate of mistakes lower than 10 %.

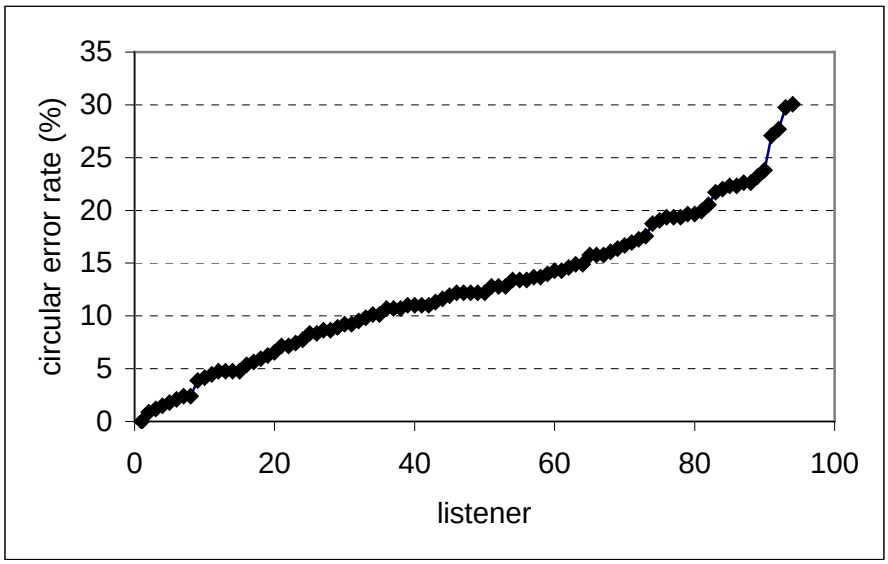

Figure 5 : Circular errors rates in the second test (loudness equalised sounds)

In this very case, it isn't possible to compute the scores of noises for listeners without taking any precaution, especially when having increasing rates of mistakes, as it was the case in the previous test. A cluster analysis revealed that the panel of listeners can be divided into 3 groups (of 35, 24 and 35 people) who have different preferences (variable preferences often appear between listeners when the effect of loudness disappears). In the following, the study of merit scores will be focused on group 3, the rates of mistakes varying between 0 and 27% in this case. This group is divided into three categories :

- in the first one (11 people), the rates of mistakes are lower than 10% ;
- in the second one (22 people), the rates of mistakes are lower than 15 % ;
- the third one represents all the people from group 3 (35 people).

A light variation can be observed in the computation of the scores of these three categories : the global difference between results diminishes. This is logical in the sense that circular errors weaken a global hierarchy : the fewest they are and the biggest is the difference between the scores of the different noises (figure 6).

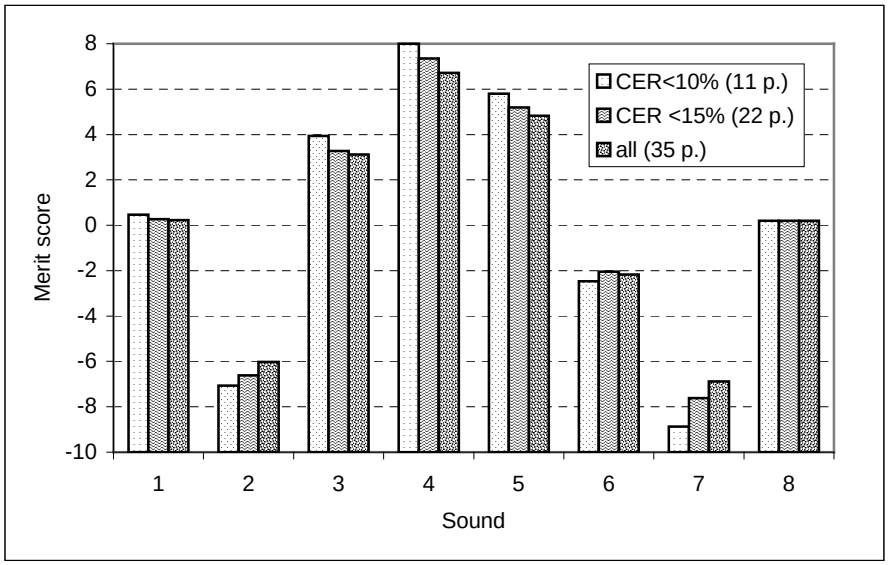

Figure 6 : Merit scores computed from differently reliable listeners results (loudness equalised sounds)

Ref : Acustica united with Acta Acustica 88 (2002), 594-598

As in the first test, a statistical study of the merit scores averaged over each group reveals no significant differences between them, excepted for sound 7, for which scores computed for the first and third group are different ($t(44)=2.43$, $p=0.05$). Therefore, it does not seem necessary to take some listeners off the panel; the exception may be the listener with the greatest circular error rate (27%).

4.3 Modification of the results

The results were modified as in the previous test according to (6) so as to simulate a test with a forced choice. The results show in this case that the coefficient of consistence of listeners are much lower than in the first test : once the different samples are determined, the probability to obtain a coefficient of consistence higher than 0.7 can be 1 for only 16 people (17% of the panel) (figure 7). In that case, a listening test with no tie allowed would have lead to reject most of the panel, though it was shown in part 4.2 that this is not necessary.

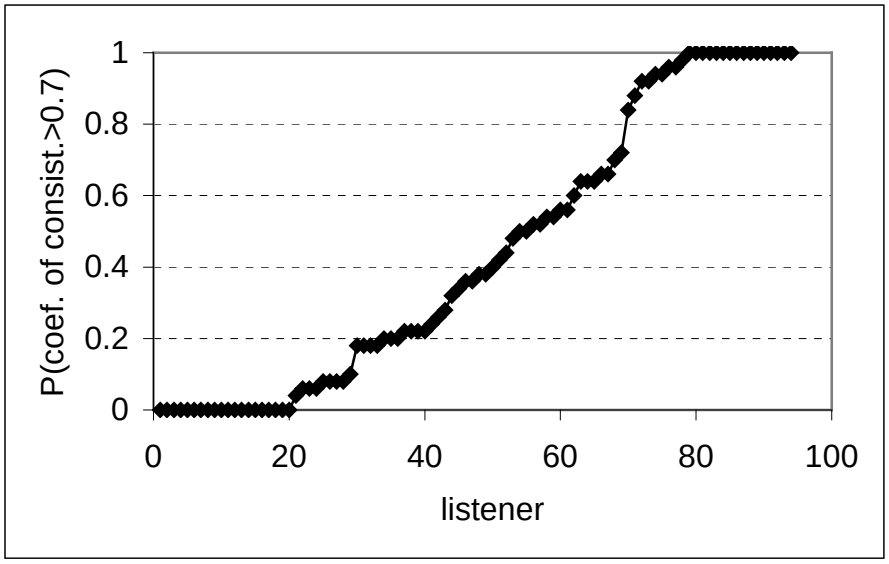

Figure 7 : Probability that $\zeta>0.7$ for different listeners (loudness equalised sounds)

5) Discussion and conclusion

The difference between the two tests is clearly due to loudness : in the first test, loudness values of the different stimuli were very different; loudness could then be used by listeners to select their preferred sounds, which made the task rather easy. Very few circular errors were committed and a forced choice test would be possible in that case. But in the second case, when loudness values were equalised, the decision was more difficult because the different stimuli were close to each other according to the evaluation which listeners had to realise; more circular errors are found. In such a case, a forced choice test would obliged a listener to select one of the two sounds of a pair while he really cannot; therefore, such a choice would be made in a random way, which increases the number of circular errors.

Johansson et al. [7] had recently insisted on the fact that paired comparisons tests with no allowed ties can give poor results if the population can be separated in two groups of people with opposite preference : for a given pair, the probability of preference can be 0.5 (half of the sample preferring the first sound and the other half preferring the second sound) and this result cannot be distinguished from the other case when no choice can be made by all listeners and the answers are random.

Clearly, another drawback of paired comparisons tests with no ties allowed is the poor results obtained when sounds are perceived as rather similar by listeners.

It can thus be recommended :

- To use procedures of pair comparison tests allowing the listener to answer that the evaluates the two noises of a pair as equally annoying This allows him not to have to choose an unsatisfactory answer at random, especially when perceptive differences between sounds are not very important ;
- To use a quantification of the number of circular mistakes (whatever the number is) as an indication of the difficulty of the test rather than a mean to systematically reject the listeners for whom the value is higher than a particular limit.

6. Acknowledgements

The author is grateful to the French railway society (SNCF), which ordered the study, and to Johan Jacquemoud, who realised the experiments.

7. Bibliography

[4] Amman, S. and Greenberg, J. "Subjective evaluation and objective quantification of automobile strut noise". Noise Control Engineering Journal, **47** (1) (1999); 17-27.

[5] Stuecklschwaiger, W and Schiffbaenker, H. and Brandl F.K. "Improving the noise quality of combustion engines". SIA Congress, Lyon (1993).

[1] Weber, R. "Interior car sound quality – assessment of acceleration noises". 2nd Forum Acusticum, Berlin (1999).

[6] Ross R.T. "Optimum orders for the presentation of pairs in the method of paired comparisons". Journal of Educational Psychology, **25** (1934); 1934.
375-382

[2] David, H.A. "The method of paired comparison". Oxford University Press, New-York (1988).

[3] Kendall, M.G. and Smith, B.B. "On the method of paired comparisons". Biometrika **31** (1940); 324-45.

[7] Johansson A-C., Hammer P. and Nilsson E. "Aspects on three methods for paired comparison listening tests". 17th International Congress of Acoustics, Rome (2001).