

Eleventh International Congress
on Sound and Vibration
5-8 July 2004 • St. Petersburg, Russia

IMPROVING THE EFFECTIVENESS OF PAIRED COMPARISON TESTS FOR AUTOMOTIVE SOUND QUALITY

Stephen Baker, Paul Jennings, Garry Dunne¹, Roger Williams²

Warwick Manufacturing Group, University of Warwick
Coventry, CV4 7AL. United Kingdom
Email address of lead author: Stephen.Baker@warwick.ac.uk

Abstract

Sounds generated by new vehicles are important for brand identification and product differentiation. Consequently automotive manufacturers are interested in how to turn customer preferences of sounds into achievable engineering targets; sound quality engineering. One tool within this process is paired comparison testing, where a jury is asked to choose between pairs of sounds given a particular question, e.g. “Which is more powerful?” or “Which is more refined?”. Such evaluations are costly and time consuming; hence artificial neural networks (ANNs) are being developing to reduce the reliance on jury evaluations.

To evaluate n sounds, $n(n-1)$ pairs are presented to the jury. This includes pairs that are presented to the jury in $i-j$ & $j-i$ orders. When preferences are averaged over jurors, pair probabilities can be calculated e.g. the probability that the sound i is preferred to sound j ($p_{i,j}$) and vice versa, which should sum to unity. However in practice due to repeatability this is not always the case as juror responses maybe influenced by the order in which the sounds are played, similarly the pair may be difficult to compare. Only one probability can be used to train the ANN.

To overcome these difficulties a paired comparison test using freeplay has been developed and assessed. In ‘freeplay’ a juror is presented with a single pair and can play and replay the sounds in any order before making their choice, thus removing the effect of pair order. Additionally the results indicate increases in jurors’ repeatability and consistency measures.

¹ W/1/009, Jaguar Engineering Centre. Abbey Road, Whitley, Coventry. CV3 4LF. United Kingdom

² Sound & Vibration Technology Ltd. Station Lane, Millbrook, Bedfordshire. MK45 2YT. United Kingdom

INTRODUCTION

Modern vehicles are becoming so quiet that sound character is now an important issue. Additionally the sounds that a vehicle makes are used by customers to distinguish between competing brands and products. Thus a significant level of effort is made to ensure that the sounds that a new vehicle makes reflect the brand image.

When a new vehicle is being developed, much work is completed using cost-effective simulation tools rather than more traditional testing and re-engineering. This is also true in the area of Sound Quality, a process that turns customers' perceptions of sounds into real engineering targets. Potential new target sounds may be generated by several methods including a full synthesis of the sound via simulation tools, to the manipulation of sounds from existing vehicles or prototypes or a combination of the two.

These target sounds are evaluated to ensure that they are inline with the brand and satisfy potential customers' expectation. The sounds should also be achievable by the correct specification of components on the vehicle. The work described in this paper concentrates on the evaluation of the sounds from the point of view of the customer's perception, rather than the achievability through component specification. In order to evaluate target sounds (or existing sounds) with respect to customers perceptions a number of methods can be used. These include the Semantic Differential test and the Paired Comparison test. Both these tests rely on recruiting a jury of existing and/or new customers.

THE PAIRED COMPARISON TEST

The paired comparison test is a simple procedure that can be used with untrained jurors (e.g. customers). A jury is asked to listen to pairs of sounds and then choose one based upon a question; "*Which sound do you think is the more powerful ?*" or "*Which sound do you think is the more refined ?*".

A statistical analysis of the responses from many jurors, compared on a pair-at-a-time basis of many sounds yields a linear scale for the quality asked in the question (powerfulness, refinement). The position of a sound on this scale is called the merit score. The merit scores are calculated from pair probabilities using Ottos' derivation [1] of the Bradley-Terry model, such that the merit scores sum to zero. This method calculates the merit score for sound i (M_i) of the n sounds compared, based on the pair probabilities p , where p_{ij} is the probability of all jurors that sound i is preferred to sound j ;

$$M_i = \frac{1}{n} \sum_{i \neq j} \ln \left(\frac{p_{ij}}{p_{ji}} \right) \quad (1)$$

These question-specific merit scores can be plotted, as in figure 1, which shows how a number of sounds can be ranked indicating the relative differences between them. These results are then used within the sound quality process. This figure shows two sets of merit scores, one from real jurors and one from a part-artificial juror, discussed later.

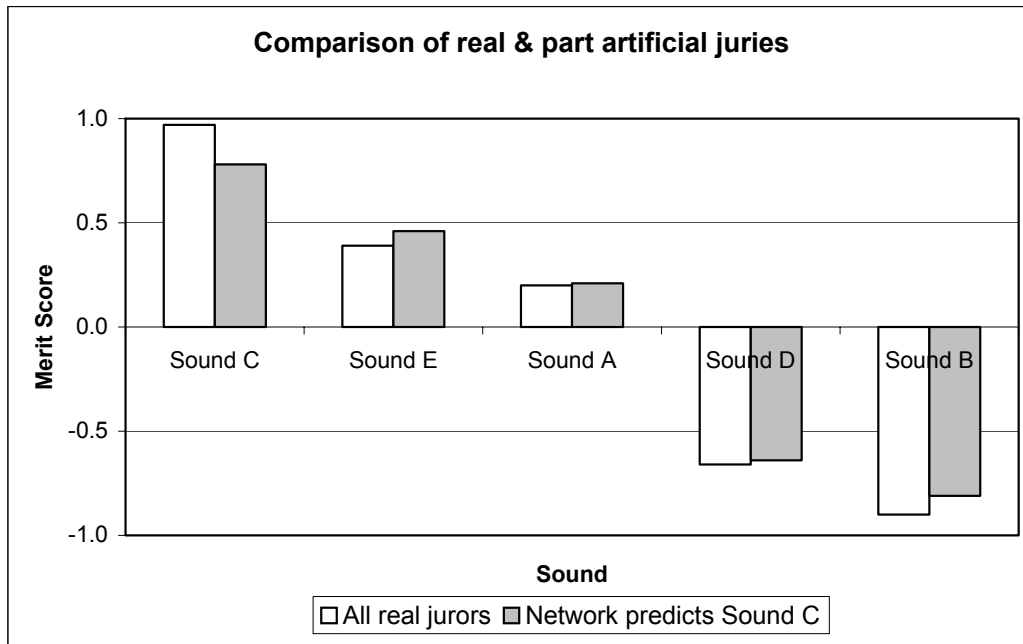


Figure 1. Merit scores from real and part artificial juries. [3]

Individual juror consistency and repeatability can also be calculated. Jurors who guess, answer at random or genuinely cannot choose consistently attain lower repeatability and consistency scores. Unreliable jurors can therefore be filtered out before calculating the merit score. The repeatability and consistency measures are discussed later in this paper.

Jury testing is a time consuming and costly exercise. For each study/sound set many jurors are required and must be recruited. Also jurors have short useful attention spans, resulting in only small sound sets being assessed. For these reasons an investigation of methods to improve or ideally replace such evaluations is being completed.

Following a previous project which made successful use of artificial neural networks (ANNs) for predicting the absolute noise level for a vehicle from its design parameters [2], it has been recognised that the use of ANNs may be used for the prediction of the pair probabilities.

Neural Network Approach

Neural networks learn predictive models from example data (training set). They are particularly useful for problems where causal links cannot be identified by traditional analytical methods. For this reason it was considered that the use of ANNs would be suitable for the prediction of pair probabilities and hence merit scores from a number of inputs.

In order to develop an ANN it is necessary to fulfil the following criteria;

- Provide the neural network with a suitable set of inputs (causes) from which it can learn the predictive model.
- Provide the neural network with sufficient known good data (training data) in order that the correct predictive model can be learnt.

Earlier work developed an ANN for predicting the pair probabilities using sound metrics as its inputs [3]. The training set consisted of several hundred jurors from the UK, USA and Germany, who compared sets of five sounds from a pool of twelve. One sound, sound C, was held back from the training set for testing the ANN.

The results from this early ANN were promising, as can be seen in figure 1. This figure shows the merit scores for five sounds (calculated from real jurors results) and those predicted by the ANN. It can be seen that the ANN correctly ranked the sounds. However to fully assess the performance of the ANN it is necessary to understand the behaviour of real jurors.

PAIR PROBABILITIES AND JUROR REPEATABILITY

During a traditional paired comparison test a jury listens to sounds presented in pairs. Firstly they listen to Sound i , and then Sound j following which the jury is asked to vote for either sound based on a question. Votes are then averaged over the jury to calculate the pair probability for the pair of sounds i, j . The probability that the sound i is preferred to sound j ($p_{i,j}$) is equal to 1 minus the probability that sound j is preferred to sound i ($p_{j,i}$). That is;

$$p_{i,j} + p_{j,i} = 1 \quad (2)$$

However in addition to listening to sounds in $i-j$ order, the jury will also listen (albeit unknown to them) to the pairs in the reverse $j-i$ order (repeat pairs) and vote again. For an evaluation of n sounds, the jury will listen to $n(n-1)$ pairs. These repeat pairs are included to see whether the play order affects the decisions made by a juror and to assess the their repeatability .

Hopefully, jurors would still vote the same way even though the order has been reversed. So, by counting the number of times the juror selects the same sound of a given pair (in both orders of play) repeatability can be calculated;

$$\text{Repeatability (\%)} = 100 \times \frac{\text{Number of comparisons judged the same}}{\text{Total number of comparisons}} \quad (3)$$

The problem lies in the situation whereby a juror does not vote the same way for the repeat pair, as this leads to an inconsistent set of pair probabilities, i.e.

$$p_{i,j} + p_{j,i} \neq 1 \quad (4)$$

This results in a number of potential pair probabilities that could be used to train an ANN. These are;

$$p_{ij} \quad \text{Or,} \quad 1 - p_{ji} \quad \text{or,} \quad \frac{p_{ij} + (1 - p_{ji})}{2} \quad (5)$$

This leads to problems when using such pair probabilities to train an ANN. The incorrect choice of pair probabilities could lead to the ANN learning the effect of play order. However if the effect of play order can be removed then this problem may be overcome as only one pair probability exists; p_{ij} .

PAIRED COMPARISON TESTING: FIXED PLAY vs. FREEPLAY

One method to remove the order of play as a variable is freeplay. Unlike a traditional paired comparison test, where the jury listening to Sound i followed by Sound j (fixed play), and then being asked to make a choice, the juror has control over the order of play and the number of times the juror listens to the sounds. In practical terms the juror has a button for each sound in the pair and may play and replay the sounds as many times as necessary and in any order until they are happy to vote.

In addition to removing the play order as a variable (and hence the potential for inconsistent pair probabilities) freeplay has a number of other advantages;

- There is no requirement to play the pairs in reverse order of play, hence the number of pairs presented to the juror is half that of a fixed play paired comparison test (i.e. $n(n-1)/2$ pairs, where n is the number of sounds in the set).
- The increased level of juror interaction may lead to greater juror retention.
- Given that a juror can replay the sounds a number of times before making a preference, it is possible that juror's responses are more repeatable and consistent. Measures of consistency are discussed later in this paper.

A study has taken place within Warwick's Listening room to evaluate the suitability of paired comparison tests using freeplay.

Twelve naïve jurors (i.e. jurors who had never completed a paired comparison test before) were recruited and split at random in to two groups. Each group completed two paired comparison tests for an identical sound set. The first group completed fixed play first followed by freeplay, and the second group conducted freeplay followed by fixed play. There was a minimum 1 day break between the tests for both groups, to prevent the jurors learning the sounds. Two groups were used to ensure that any differences in the results was due to the difference in test type rather than the order (or number) of tests completed. The test details were;

The sounds. The sound set used was 6 sounds of cars accelerating under second gear wide open throttle (2GWOT) conditions, with the sounds trimmed to include engine speeds from 2500 rpm upwards. This results in 15 *i-j* order and 15 *j-i* order (repeat) pairs.

Fixed play test. All pairs were presented to the jury, hence 30 pairs in total. The same 30 pairs were presented once for the question “*Which do you think is the more powerful sound ?*” and once for the question “*Which do you think is the more refined sound ?*”.

Freeplay test. Since play order is not a variable only 15 pairs were presented to the jurors. However, 10 pairs were repeated (selected at random) resulting in a total of 25 pairs of sounds being used. Again these 25 pairs were used twice for the evaluation of the powerful and refined questions.

Jurors also completed a short questionnaire after each test. The questionnaire was designed to elicit how involved the juror felt in the test; the ease with which they could discriminate between the sounds; and if there were differences between the powerful and refined questions.

The results of the two test methods were compared on the basis of individual juror’s repeatability and consistency measures.

Juror consistency can be indicative of the reliability of their responses. Juror consistency is measured by the coefficient of consistency [4], which is based on the concept of circular triads.

Consider three sounds A, B and C. If a juror prefers A to B, and B to C, it follows that the juror should prefer A to C. If this is not the case then the triad is said to be circular, in that no dominant sound is preferred. By counting the number of consistent triads (*v*) over the number of possible triads (*m*) a value of consistency (ξ) can be calculated;

$$\xi (\%) = 100 \times \left(\frac{v}{m} \right) \quad (6)$$

RESULTS

A summary of the results is shown in table 1, which lists the juror ID and the order in which the juror completed the tests. The Δ values shown are the absolute differences in the consistency and repeatability measures (which are percentages) between freeplay and fixed play for both the powerful and refined questions. e.g. Δ consistency is a jurors freeplay, powerful consistency score (%) *minus* the same jurors fixed play, powerful, consistency score (%), resulting in a percentage difference ($\Delta\%$).

As such a negative Δ value indicates that the jurors' performance was better during the fixed play test, and a positive Δ value indicates greater performance during a freeplay test. For example; juror 34s powerful consistency was 4.3% better in fixed play than freeplay, however their refined consistency score increased by 6.7% from fixed play to freeplay.

		Freeplay <i>minus</i> Fixed Play			
		Consistency ($\Delta\%$)		Repeatability ($\Delta\%$)	
ID	Test Order	Powerful	Refined	Powerful	Refined
34	Freeplay - Fixed Play	-4.3	6.7	10.0	0.0
35	Freeplay - Fixed Play	2.5	-3.3	0.0	13.3
36	Freeplay - Fixed Play	-0.8	-4.5	10.0	-23.3
37	Freeplay - Fixed Play	8.7	0.7	23.3	13.3
38	Freeplay - Fixed Play	-1.4	10.7	-16.7	20.0
48	Freeplay - Fixed Play	-7.7	1.0	-3.3	3.3
40	Fixed Play - Freeplay	6.2	13.7	33.3	36.7
41	Fixed Play - Freeplay	12.5	21.8	26.7	26.7
42	Fixed Play - Freeplay	4.4	21.0	16.7	-10.0
43	Fixed Play - Freeplay	-2.3	25.0	6.7	23.3
44	Fixed Play - Freeplay	12.5	3.7	33.3	26.7
49	Fixed Play - Freeplay	1.6	10.7	16.7	16.7

Table 1. Differences in repeatability and consistency between freeplay & fixed play.

DISCUSSION

The results indicate that test order has an effect. It can be seen that almost all the Δ values (consistency & repeatability for powerful & refined) increase when moving from fixed play to freeplay, whereas moving from freeplay to fixed play a fewer Δ s increase. This may show that the (naïve) jurors are learning the test and in particular defining to themselves what they determine a powerful or refined sound. This is supported by some of the questionnaire response. For example one juror states “*I kept asking myself, what is refined ?*” during the first test they completed. Another juror, when asked to rate their own consistency in their given answers, stated “*I was consistent once I had decided on a criterion*”.

That being said, it is clear that the results generally show greater consistency & repeatability measures for freeplay compared to fixed play, irrespective of the order of the tests.

Although not shown here if the mean values for consistency & repeatability for powerful & refined (when averaged across jurors for each group) are calculated, it can be seen that for those who completed the freeplay test first demonstrate greater consistency by up to 2% (for powerful and refined), and similarly a greater repeatability by up to 4%. For those who completed the freeplay test second, the gain is even larger with consistency and repeatability measures increasing by between 5-16% & 20-22% respectively.

Whilst these results are encouraging, a more thorough analysis of these figures is needed. In order to assess any potential increase it must be done with respect to variability between jurors, test, and the powerful & refined questions. This will give a more accurate measure of any increase and a level of significance to be attributed to any change. Further studies are required with more jurors (both naïve and experienced), to provide a larger data set for a thorough statistical analysis. This is currently underway.

CONCLUSIONS

The results presented indicate that a juror's repeatability and consistency measures are improved by the use of freeplay within a paired comparison test by up to 20%. This is due to the nature of freeplay, not the experience of the juror. This may mean that fewer jurors are recruited to find the required number with acceptable repeatability and consistency. Increased consistency will result in a greater confidence in the merits scores. Juror retention may also be improved: questionnaire responses indicated that all jurors felt more involved in the freeplay test.

ACKNOWLEDGEMENTS

The work described in this paper was funded by the UK Engineering and Physical Research Council (EPSRC), through the Warwick Innovative Manufacturing Research Centre.

REFERENCES

- [1] N. Otto, *Listening Test Methods for Automotive Sound Quality*, 103rd Audio Engineering Society Convention, 1997
- [2] J. Fry, P. Jennings, N. Taylor & P. Jackson, *Vehicle Drive-By Noise Prediction: A Neural Networks Approach*, SAE Transactions 1999, Volume 108, Section 6 - Journal of Passenger Cars, Part 2, pp 2769 - 2775, September 2000. ISBN 0-7680-0697-X
- [3] P. Jennings, J. Fry, G. Dunne & R. Williams, *Predicting Customers' Evaluations of New Vehicle Sounds*, Proceedings of the 10th International Congress on Sound and Vibration, 7-10 July 2003. Stockholm, Sweden.
- [4] MTS Jury Evaluation Version 2.0 software manual. MTS Software Systems.