

Understanding how customers make their decisions on product sound quality

Jeff Fry^a, Paul Jennings^b, Roger Williams^c and Garry Dunne^d

^{a,b}Warwick Manufacturing Group, University Of Warwick, Coventry. CV4 7AL United Kingdom.

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

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

^{a,b}[jeff.fry;paul.jennings]@warwick.ac.uk;

^crwilliams@sovibtech.com; ^dgdunne@jaguar.com

Abstract 170 A product's sound can reinforce its brand image. Sound quality engineering is concerned with turning customer preferences, brand attributes and marketing information into product engineering targets. Automotive companies for example are interested in how their new car should sound, since sound is an important way to differentiate products. Currently, candidate sounds are evaluated by costly and time consuming jury evaluations, often using the method of paired comparisons. Neural networks have been identified as a method of evaluating a large number of sounds quickly and cheaply. A Multi-Layer Perceptron has been trained from existing jury data to predict the probability of a juror selecting one sound of a pair. These results can simulate a jury evaluation and give rankings of new and existing sounds. Neural network inputs include objective sound metrics - numeric or categorical inputs that describe the sounds. However, the difficulty is to decide which of the numerous, time-varying metrics to use and how to represent them as single numbers. Further evidence is required to choose the optimum inputs to present to the neural network. One option is to choose the inputs based on an understanding of how customers actually make their decisions. To this end, a simple interview procedure has been developed and tested. This is described and sample results are presented and assessed. The suitability and wider application of this procedure is discussed, and further work suggested.

1 INTRODUCTION

The various sounds made by a product convey information to its user about the product quality and performance. These impressions can help to reinforce the brand image. Ideally therefore, sound quality is considered and engineered into products at an early stage. Sound quality engineering is concerned with turning customer preferences, brand attributes and marketing information into engineering targets. Automotive companies are particularly interested in sound quality engineering, since sound is an important way that customers differentiate cars [1]. Door closure sounds for example, give an impression of the quality and reliability of engineering in a vehicle. Engine sounds inform a driver of the refinement and performance of the vehicle. The difficulty for manufacturers lies in specifying what target sound a new vehicle should have [2]. Potential targets can be created

by manipulating or mixing existing sounds, or by simulation. In reality, there will be many constraints on the vehicle sound, such as cost, weight and manufacturability that require compromises. Therefore, during the design process, the target sound may be revised. New targets may need to be evaluated for customer acceptance, but sadly, company experts are not usually able to predict customers' reactions accurately. Consequently, sounds have to be evaluated by costly and time consuming jury evaluations, often using the method of paired comparisons.

The paired comparison method is a simple task, appropriate for non-experts, who are presented with a pair of stimuli, and asked to choose one in response to an evaluation question. For example, a car manufacturer may present their new, existing and competitors' car sounds and ask the evaluation question, "which sound is more powerful?". A quicker, cheaper, yet robust method of evaluating sounds automatically would be of considerable use. Such a method could initially filter sounds, leaving the better candidates to be evaluated by real customers. Previous work has identified neural networks as a potentially suitable method to evaluate sounds [3]. To date, two evaluation questions, about the powerfulness and refinement of sounds have been asked. Multi-Layer Perceptrons (MLPs) have been trained from existing jury data to predict the probability of a juror selecting one sound of a pair [4], for each of these two questions. The pair probabilities are learnt from existent jury evaluation data. The pair probability predictions for all pairings of sounds from a group can be assembled to simulate the results obtained from jury evaluations. These results are analysed using the Bradley-Terry model, described by David [5] to rank the new and existing sounds. Merit scores place the sounds on an arbitrary linear scale, particular to the evaluation question and unique to the set of sounds evaluated.

The inputs to the MLPs are numeric values derived from the objective sound metrics that describe the sounds. However, one difficulty is to decide which of the myriad, available time-varying metrics to use and how to represent them as single numbers. Tenth or ninetieth percentiles, minima, averages, maxima, average gradient and standard deviation of the metric curves are all possible treatments. Hitherto, inputs were either suggested by experts, or discovered by multiple linear regression of merit scores against metrics. Although the outcomes were encouraging, the choice of inputs has been somewhat arbitrary, and the regression method gives answers particular to the set of sounds used in the evaluation. Further evidence is required to choose an optimum and generalised set of inputs for the powerful and refined MLPs. A potential and principled solution is to choose inputs based on an understanding of how real jurors actually make their choices in paired comparison evaluations. To this end, a simple juror evaluation and interview procedure, lasting less than one hour has been developed and tested.

The procedure was developed through successive trial and refinement, initially with jurors from the first authors' organisation. After a number of iterations, the procedure was used to collect results from jurors working at all the authors' organisations. These jurors represent two distinct populations, noise, vibration and harshness experts and non-experts. Eventually the procedure will be used with real luxury car customers. This paper describes the procedure and suggests a method to analyse the results. Initial outcomes are presented, along with a discussion of some of the issues encountered. Note that it is not the authors' intention to undertake a rigorous psychology experiment, but rather to devise a practical method to examine how people make choices during a paired comparison evaluation of automotive sounds, which may have wider utility.

2 JUROR EVALUATION AND INTERVIEW PROCEDURE

The procedure is based on the traditional paired comparison evaluation technique with a simultaneous interview. This is in order to capture jurors' thought processes during a paired comparison evaluation as closely as possible. It is implemented as a PowerPoint slide show running on a laptop PC. The interviewer records the juror's choices and answers on a simple paper form. It is also recorded, with the juror's permission, on a Digital Voice Recorder. It all takes place in an acoustically isolated Listening Room, and follows established guidelines for listening tests [6].

2.1 Procedure design

The procedure consisted of an introduction, two paired comparison tasks, one sound description task and a de-briefing. These tasks were primarily intended to find which sound features jurors used to distinguish between sounds. The progression of the slides was under the juror's control and presents the tasks and associated questions. The two paired comparison tasks were standard forced choice evaluations, (no ties are allowed) for the powerful and refined questions. The juror could play each sound of a pair as often as they need, and in any order – referred to here as “free-play mode”. This contrasts to the more common practice, referred to as “jury mode”, where perhaps six jurors simultaneously evaluate pairs, and each sound of a pair is played once only before they vote. The interviewer noted the reasons given by jurors for their choices, using the jurors' own words, on a paper form.

Additionally, the juror was also asked to rate how difficult it is to choose between each pair for both evaluation questions. This allowed the experimenter to quickly identify which pairs are the easiest to evaluate. Easy pairs are often the least interesting from the point of view of gathering reasons used by jurors in their choices. Generally, most jurors reported the same few features, and once those reasons have been found, there is little benefit asking more jurors to evaluate the same pairs. It is often the difficult pairs that yield the more interesting information. New pairs can then be put in place of the easy pairs, to improve the efficiency of data collection.

The meanings of words used by jurors in the evaluation tasks was checked in the subsequent sound description task, where each sound was presented on its own. The interviewer was careful to ask non-leading and open-ended subsidiary questions where clarification was required. Prior practice by the interviewer is therefore recommended. Finally, in the de-briefing, the juror was asked to describe the strategies employed to make their choices for each of the evaluation questions. This was to try to discover any general rules used by jurors when making their choices. Juror feedback was also sought at this stage to help improve the procedure itself.

2.2 Procedure structure

The actual procedure used to gather data in this work can be summarised as follows:-

Introduction

- the juror completes a demographic form and is reassured about privacy and data protection
- the juror is briefed on their evaluation
- the juror is given a context using exterior and interior photos and the example sounds
- the juror is introduced to the paired comparison task using two practice pairs of example sounds
- permission is sought to record the interview

Paired comparison tasks

- for each evaluation question (powerful, then refined)

- the juror is reminded there are no trick questions and no “right” or “wrong” answers
- for each pair of sounds
 - the juror chooses one sound of the pair
 - the juror is asked to explain how they made their choice
 - the juror is asked how difficult the choice was to make

Sound description task

- for each sound
 - the juror is asked to describe the sound
 - the juror is asked to rate how powerful and refined the sound is

De-briefing

- the juror is asked for feedback about the interview and their previous experience
- the juror is asked which evaluation question was more difficult
- for each evaluation question (powerful, then refined)
 - the juror is asked to describe their strategy to choose between the sounds in a pair
- the juror is asked for any other comments or observations
- the juror is asked to confirm whether the digital voice recording can be kept.

2.3 Analysis method

The data gained from this procedure is mostly textual, rather than numeric so it is worth describing the analysis method results in some detail. It has the following stages: -

1. transcribe interviews
2. split explanations into individual ‘reasons’
3. express each reason in a standard form
4. assign a keyword to each reason
5. resolve opposite reasons
6. group similar reasons.

There are many psychology textbooks covering questionnaire design and qualitative data analysis, Coolican [7] for example, gave a helpful overview.

Each juror interview was transcribed from the paper forms and the digital voice recordings. The explanations, given by jurors for their choices in the powerful and refined paired comparison evaluations, were split into one or more individual ‘reasons’ and labelled appropriately to allow tracing back to the original words. If a juror gave a reason more than once, for a particular pair, the repetitions were discarded. Each reason was represented in a standard form, as follows...

{Expert|Non-expert} juryID “says” soundX “is more {powerful|refined} than” soundY {because|despite} {soundX|soundY} reason **keyword** [at start|in middle|at end|throughout]

In this form, curly braces {} mean select one of the enclosed items and square braces [] mean select 0, 1 or more of the enclosed items. Items are separated by |. juryID, soundX, soundY and reason are derived from the juror’s answers. Literal text is enclosed in double quotes. Reasons may pertain during one or more time-periods of the sounds. Keywords (or key phrases) should be carefully chosen to capture the gist of the explanations given.

This scheme allowed quite complex reasons - both positive and negative - to be stored in an Excel spreadsheet which aided further processing. It also captured information about which time-periods of the sound jurors find important. An example may clarify how one juror’s answer to the powerful

question for the pair (X, Y) was processed. The interviewer noted the answer: “X accelerates very quickly. Y has engine modulated growl at start. X is more powerful, but I prefer Y”. This was split into four statements, using supporting evidence from the voice recording as follows:-

Expert 89 says X more powerful than Y because X accelerates very quickly **acceleration rate**

Expert 89 says X more powerful than Y despite Y having engine modulated growl **growl** at start

Expert 89 says X more powerful than Y because X is more powerful **powerful**

Expert 89 says X more powerful than Y despite Y being preferred **preference**

The statements were then sorted by keywords, so that similar reasons can be grouped and counted. Keyword opposites (e.g. ‘loud’ and ‘quiet’) and near-synonyms (e.g. ‘irritating’ and ‘annoying’) were then resolved to amalgamate groups where possible. Subtle distinctions were noted that allowed groups to be sub-divided, for example, jurors readily distinguished between ‘acceleration rate’ and an ‘even acceleration’. The statements referring to ‘powerful’ were really redundant. The method identifies the most common reasons, counted over all jurors and all pairs of sounds.

3 RESULTS

The procedure was used to evaluate seven pairs of sounds for powerful and refined. Ten jurors were experts from Jaguar and Sound & Vibration Technology, and ten, non-experts from Warwick Manufacturing Group. Sixteen jurors had taken part previously in a jury evaluation, four of the non-experts had not. The jurors are not claimed to be representative of customers, but they can still help to develop the interview technique, and reveal if experts and non-experts give different reasons for their choices.

Sounds from nine luxury sports cars, accelerating in 2nd gear at wide-open throttle (2GWOT), were recorded in-vehicle on a binaural acoustic head. These vehicles included two Jaguar models and their main competitors. Seven of the nine sounds were evaluated by paired comparison; the others were used to demonstrate the tasks required of the juror. Seven sound pairs were predefined, and each sound appeared in two pairs. The sounds were replayed at their correct loudness level through a Digigram VX Pocket v2 24-bit digital PCMCIA sound card, Stax SR 3030 headphone amplifier and Stax SR 303 Classic electrostatic open headphones. Interviews were conducted in equivalent Listening Rooms at the University of Warwick and Jaguar’s Engineering Centre.

3.1 Reasons for powerful choices

There were 469 individual reasons expressed in stage 3 of the powerful reason analysis. In stage 4, one hundred distinct keywords suggested by the reasons themselves, were assigned. Table 1 ranks in descending order, the forty five most frequently assigned keywords, which together account for over 84% of the reasons given (and the count is ≥ 3). The results from this stage are presented, so the reader can appreciate the subjective nature and scale of the final stages, and perhaps make their own conclusions. Notice that in this work, only one third of the possible pairs were evaluated. For completeness the remaining pairs may be evaluated at another time.

All the keywords have yet to be resolved, grouped and ultimately translated into metrics and treatments. Even at this stage of the analysis, some interesting points emerge. ‘Acceleration rate’ and ‘loudness’ were, not surprisingly, very popular keywords when thinking about the idea of powerful. ‘Powerful’ was of course, redundant. ‘Whine’, which could probably be grouped with the less popular ‘supercharger whine’ and ‘supercharger noise’ was nearly always mentioned in a negative sense. Similarly, ‘low frequency content’ was nearly always a positive feature.

Rank	Keyword	Count	Rank	Keyword	Count
1	acceleration rate	53	22	grunt	5
2	loudness	34		modulation	5
3	powerful	29		refinement	5
4	whine	28		revviness	5
5	low frequency content	14		thrashiness	5
6	growl	13		weakness	5
7	duration	12	30	even acceleration	4
	smoothness	12		muted	4
	speed	12		noise	4
10	engine sound	11		resonance	4
11	spectral balance	9		tinniness	4
12	character	8	35	aggression	3
	orders	8		annoyance	3
14	bassiness	7		no of cylinders	3
	closeness	7		definition	3
	effort	7		direction of travel	3
	muffled	7		low order content	3
	preference	7		pleasance	3
	19	blandness		6	raspiness
19	harshness	6		subdued	3
	roughness	6		throatiness	3
	22	boominess	5	wildness	3
	constancy	5			

Table 1 *Top keywords assigned to jurors' reasons*

3.2 Juror feedback about the procedure

The juror evaluation and interview procedure was developed and improved through juror feedback. Interviews lasted between forty and sixty minutes depending on the loquacity and enthusiasm of the juror. The juror feedback suggested that the procedure structure and duration are reasonable. The first author conducted all twenty interviews, and recommends a half-hour break between interviews, attempting a maximum of five in a day. The tasks were not found difficult, and jurors said they were explained well by the slides. Anticipated concerns about using a voice recorder proved unfounded. Anecdotal evidence suggested that jurors like free-play, and find the experience positive in some way. It is to be hoped that similar positive responses will be obtained when members of the public are interviewed.

Jurors' comments about the procedure are summarised below:-

- Twelve jurors thought the duration acceptable, including one who was prepared to take longer
- Only one expert thought it too long; their interview was equal longest.
- Three experts found the paired comparison task difficult, whereas only one non-expert did.
- One expert reported that the sound description task was difficult, whereas two non-experts did.
- Two non-experts were concerned their choices might change if they took part again.
- Nine non-experts found the powerful evaluation easier than refined.
- Five experts found powerful easier, three found refined easier, the other two found both easy.
- No one wanted to review or delete his or her voice recording.

- Three non-experts and one expert explicitly stated their preference for free-play over jury mode.
- None of the sixteen jurors who had taken part before in a jury evaluation explicitly preferred jury mode to free-play.
- Five non-experts and three experts made positive comments about the experience, two saying that they had fun.
- One non-expert and five experts thought it challenging or interesting
- Three non-experts and four experts considered the procedure useful in some way.

4 AGENDA FOR FURTHER WORK

Several issues with the procedure and analysis method remain to be resolved. Firstly, the choices made by jurors should be compared to similar results obtained without the juror interview, to decide whether the procedure changes the jurors' answers. If the choices can be shown to be consistent, then the procedure outcomes can be assumed to be representative.

Secondly, the iterative grouping stage of the analysis, stage 6 is, to some extent, subjective. The analyst should be careful to be guided by the text, rather than by preconceived theories about how jurors make their choices, or a desire to justify use of a particular metric. To reduce the subjectivity of the analysis, it would be worth asking several independent people to categorise the reasons and reconcile their results. There are several lexicons of terms used in automotive sound quality being developed [8] & [9] with a view to producing a standard. Choosing keywords from a standard lexicon would greatly benefit the analysis if each lexicon word could be correlated to sound metrics once, in a widely accepted manner. The lexicon would then be a valuable intermediate step in subjective to objective translations.

The procedure has been developed with experts and non-experts that do not represent real luxury car customers. The next important step is sample from this population. The authors' colleagues were prepared to donate an hour for this work, but recruiting from the public is expected to be harder. They will have less motivation to help, and their overheads, in terms of time and effort to attend will be much greater. To this end, a portable listening room is envisaged, that could be taken to where customers already are; car showrooms, car enthusiast meetings and visitors to Jaguar factories perhaps. Some work on delivering paired comparison evaluations on the Internet has also been carried out. Non-standard, uncontrolled listening environments, and juror reliability cause concerns, but there may be an acceptable trade-off between the data quality and quantity obtained.

Finally, the reasons discovered in the analysis should be used to select treated metrics to use as MLP inputs. One way to accomplish this is to examine the reasons on a pair by pair basis and try to match the keywords to those metrics that show the greatest difference for the two sounds. These links would still be subjective, but at least would be based on juror evidence. Subjectivity could be managed somewhat by a poll of experts or other means.

Both the reasons and the time-period information suggest treatments for the relevant metrics. For example 'acceleration rate' and 'even acceleration' suggest the average gradient and the standard deviation of the rpm vs. time curve respectively. Also, if, for example, 'at start' is frequently reported for one metric, that part of the curve can be given special consideration. Ultimately, the metrics and treatments found by this procedure, would be used to train several MLPs, each having different sub-sets of inputs. Comparing the performances of the MLPs would then identify the most successful set of inputs.

4.1 Wider application

This procedure may be of interest to manufacturers of products where subjective evaluation by customers is costly, and automation would be an advantage. It could be adapted for:-

- other subjective questions, such as suitability of a sound, or sound preference,
- other sounds, such as door closures, seat adjuster motors, switch indicators etc.
- other products where sound quality may be important, e.g. power tools, white goods etc.

5 SUMMARY

Neural networks are useful for the efficient evaluation of target sounds of new vehicles. Evidence is required to decide which objective sound metrics should be used as inputs to the neural network. An understanding of how customers make choices about product sound quality can help identify the metrics and how they should be treated. An interview technique to discover how jurors choose in paired comparison evaluations has been described. It was developed using juror feedback and used in powerful and refined evaluations of 2GWOT luxury sports car sounds.

The powerful data from twenty jurors has been analysed to determine which reasons were most frequently given for jurors' choices. For the perception of power in luxury sports cars, the most important reasons include 'acceleration rate', 'loudness', 'whine' and 'low frequency content'. It remains for the reasons to be translated into a selection of treated sound metrics. Once that has been achieved, data should be collected from a new sample more representative of potential luxury car customers. A deeper analysis of these and further interviews may also help explain the differences between expert and customer opinion.

ACKNOWLEDGEMENT

This work described above was funded by the United Kingdom Engineering and Physical Sciences Research Council (EPSRC), through the Warwick Innovative Manufacturing Research Centre.

REFERENCES

- [1] Richard Lyon, *Product Sound Quality, Goals and Methods*, Keynote Address, SAE Conference on Automotive Noise, Vibration and Harshness, Traverse City, Michigan, May 1999
- [2] Dunne G, Wheeler A & Jennings P, *The Identification of Powertrain Sound Quality Target Sounds*, August 2000. Proceedings of Inter-Noise 2000 August 2000 ISBN 2-9515619-6-2
- [3] Paul Jennings, Jeff Fry, Garry Dunne & Roger 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] Jeff Fry, Paul Jennings, Garry Dunne & Roger Williams, *Jury Evaluation of Sound Quality Using Neural Networks*, December 2002. Proceedings of the 4th International Conference on Recent Advances In Soft Computing, Nottingham, UK, 12th - 13th December 2002, pp 101-102. ISBN 1-84233-0764
- [5] David, H A, *The Method of Paired Comparisons*, 2nd Edition, 1988. Charles Griffin & Co. Ltd., London UK. ISBN 0-85264-290-3.
- [6] Otto N, *Listening Test Methods for Automotive Sound Quality*, 1997, 103rd Audio Engineering Society Convention.
- [7] Coolican H, *Research Methods and Statistics in Psychology*, 3rd Edition, 1999. Hodder & Stoughton Ltd., London UK. ISBN 0 340 74760 9.
- [8] Lyon R, *Product Sound Quality, From Perception To Design*, Sound & Vibration, 2003, pp 18 – 22.
- [9] Eaton W C & Cville G V, *Lexicon For Product Sound Quality*, Canadian Acoustics – Acoustique Canadienne vol. 25, no. 3, Sept 1997, p32.