Audio Engineering Society
Convention Paper
Presented at the 110th Convention
2001 May 12–15 Amsterdam, The Netherlands

IPA – A Subjective Assessment Method of Sound Quality of Car Sound Systems

Emanuele Ugolotti [(1)], Gino Gobbi [(1)], Angelo Farina [(2)]
[(1)] ASK Automotive Industries, via Fratelli Cervi n. 79, 42100 Reggio Emilia - Italy
[(2)] Industrial Eng. Dept., Universitàdi Parma, Via delle Scienze - 43100 Parma - Italy

ABSTRACT

The paper describes a new subjective evaluation method of the acoustical quality produced by a sound system inside a car compartment. The method produces a single rating number, called IPA (Index of Performance Acoustic), defined as a weighted average of the subjective responses to a questionnaire, being compiled during listening tests conducted with the subject seating inside different cars.

The paper describes the details of the subjective test, and focuses on the choice of questions in the questionnaire and of the weight to be employed. The principal innovation of the new method is the fact that the weights are changed according to the reliability of the subject (which is also inferred from the questionnaires), and thus the evaluation is very robust and almost immune from the inclusion in the panel of completely unreliable evaluators.

INTRODUCTION

Although many advanced measurement techniques [1,2] have recently been developed for the objective characterization of the acoustical response of sound systems installed in small compartments (i.e. cars), the results of the experiments are usually not easily correlated with the subjective performance. On the other hand, it was attempted to substitute traditional listening tests inside cars with some sort of virtual presentation of the sound field through auralization techniques [3,4,5].

But the final decision regarding the choice of the sound system to be installed in a new model of car, is actually done by the car manufacturer with the traditional method of comparative evaluation of some concurrent sound systems, installed on identical cars. And it can be foreseen that this procedure will remain the same for many years to come.

The evaluation is usually simple-blind (the listeners can see and touch the different equipment, but they do not know the brand-name of the manufacturer), and they are left free to listen at will the soundtracks they like on each car.

It resulted advisable to develop a new software tool for the automatic processing of the responses collected in such comparative tests, which were previously easily biased by a weighting technique based more on the job title of each listener than on his skillness and reliability.

The software tool was just a part of a complete formal definition of the conditions to be kept during the tests, which included the prohibition to comment the listening experience together with the other listeners, the choice and calibration of the opposite-attributes subjective questions to be included in the questionnaire, and the rules for filling up the questionnaires (the order of the cars was shuffled randomly, and other not-acoustical bias was kept to the minimum reasonably obtainable for simple-blind tests, by ensuring that all the cars are absolutely identical except for the different sound systems).

The choice of optimal weighting functions is based on two concepts. First of all, a different relative weight is given to each question of the

questionnaire, according to previous subjective research [3,6] which made it possible to evaluate the relative importance of different perceptual factors on the overall quality judgment.
Then, the responses of each subject are scaled according to a second weighting factor, which depends on the degree of coherence of the responses expressed by each subject. The coherence is estimated by applying a psychoacoustics model to the subjective responses, which evaluates the degree of reliability of the subject. This means that subjects who give contradictory results are given a little weight on the overall weighted average, whilst subjects giving perfectly consistent response receive much higher weight.
The weighted responses of all the questions and all the subjects for each given car are summed in a single-number score, which is scaled to the interval 1..10, and is assumed to correspond globally to the sound quality of the system. This single-number evaluator was called IPA (Index of Performance Acoustic).
The results of some subjective experiments are reported in the paper, showing how the new method can produce discrimination between sound systems with very similar performance, provided that a reasonable number of sharped-ear subjects are included in the judging panel. When this is done, even including an equal number of absolutely unreliable subjects (as usually top managers of car manufacturing companies are), the statistical results do not suffer of any relevant bias, because the questionnaires of these unreliable subjects are automatically weighted-out by the algorithm.
Thanks to these appealing properties, the IPA score is now being employed by some European car manufacturers, and it could be proposed as a common procedure for the others.
The paper includes the sets of questions and weighting functions which were refined till now, based on listening tests conducted with South-Europe car manufacturers. It remains to be investigated if these sets of questions and weighting functions are substantially general, or perhaps they need adaptation for different car markets (Far East, South America, etc.): this will be the goal of a further research, which will be conducted in Brazil during 2001.

METHODOLOGY

In evaluating an event characterized by strong subjective aspects, one can not proceed exclusively with deterministic methods, except for exact knowledge of the phenomenon from a physiological point of view: in effect, we find attempts of this type in the literature, with the construction of the so-called “psycho-kinetic models” that approximately simulate human psycho-acoustic behavior.
Nevertheless, as of today, no one has successfully developed a model that takes individuality into account, which is to say, the preference for one sound over another, and most of all, the preference for a given sound landscape over another.
We are still lacking too much data to be able to construct an algorithm, no matter how complex, that is capable of providing an exhaustive description of the psycho-acoustic impact of sound reproduction.
It, therefore, becomes necessary to take a statistical approach.
Let's take a look at how this is possible, using a limited number of descriptors associated with an even such as sound reproduction in any environment, such as the passenger compartment of a car in this specific case.
In general, to describe any physical event, we are used to employing particular descriptors that derive from an analysis, even though approximate, using typically mathematical methods that we are more or less familiar with. This also occurs for the phenomenon of sound: for example, we are used to identifying a distribution of the Pressure scalar as a function of the frequency by applying the powerful transformation tool of Fourier and Laplace. But our experience is that, if on the one hand we have perfectly identified and quantified the problem, on the other hand, we still find it very complicated to qualify it. Or rather, a qualification is made, but only on the basis of past, empirical experience, based only on the historical memory acquired over the course of years by those who work directly in the sector: a sort of association by successive similarities. It is equally clear that such a method can lead to rough evaluations and, at any rate, not without polarization, which is to say, systematic errors that are difficult to extirpate. In addition, you don't have a real, or overall, evaluation, but rather of one point: to put it in terms of physics, knowing the motion of one particle, even perfectly, does not imply knowing anything about the gas of which it is a part!
Let us reason, then, in statistical terms and introduce the attributes necessary for extracting the information for qualifying the phenomenon of sound in a coherent and realistic way. These attributes are, in their turn, amply dealt with and documented in the literature. Looking at them, they are:

- Spatiality
- Sound Level
- Quality of Piercing Notes
- Voice Quality
- Bass Quality
- Pleasantness

The statistics that are constructed on the basis of the interpretation of these attributes are provided by a certain number of listeners. Each listener is requested to provide, in addition, to their own generalities, their own personal information about the above-cited attributes relative to the system that they are evaluating. All these data are entered in the software which was designed for automating the processing of the subjective response. The first screen of this software is shown in fig. 1.

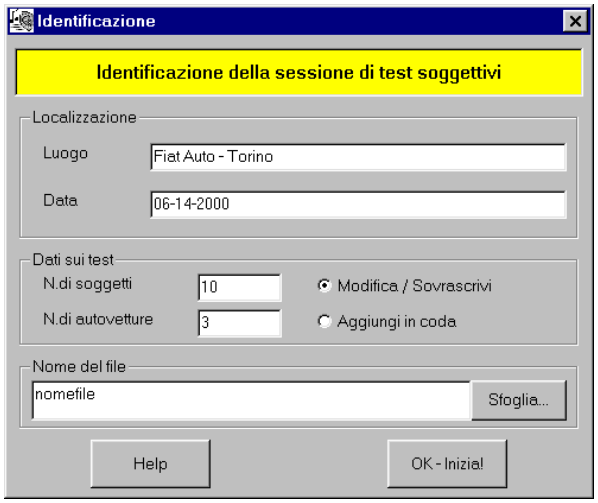

Fig. 1

To construct statistics that allow for subsequent processing, we need numeric data in order to arrive at an unequivocal judgment scale, a score, that clearly and cleanly qualifies the performance of the acoustic system: the IPA.
We then need to define the laws for assigning the subjective interpretation of the attributes, as well as the relative weight of each individual attribute: but this is simple. It's sufficient to refer to any evaluation scale: for our comfort, and for reasons of standardization, we refer to the SAE scale. It consists of a series of numbers, all between 1 and 10, and usually the lowest number is associated with the worst possible condition, while the highest number is associated with the best possible condition. Every listener is asked to associate a number on this scale that they feel is appropriate to each attribute. So as not to create confusion and, at the same time, not require an excessively difficult decision, we chose five numbers from within the SAE scale, specifically the odd numbers: 1, 3, 5, 7, and 9. As regards the relative weights, we have to keep in mind that their sum must always add up to one: as we see, they can be “static”, constant, and/or “dynamic”, which means variable.
We set up a pre-printed form that is given to each listener to fill out: at the beginning, before listening to the system, the listener writes his

own personal information, the date and the vehicle he is evaluating. Only afterwards the subject enters in the vehicle, get comfortable first in the forward listening position and then in the back, and by operating a CD or tape player, he begins listening to a pre-established musical program. During, or immediately after listening, he can write his judgments on the paper form, the design of which is shown in figure 2, filling in one of the five circles for each attribute. We point out that the best and worst condition is shown for each attribute. Each circle is associated to one of the five odd numbers 1, 3, 5, 7, 9 and the lowest number is associated with the circle closest to the worst condition, while the higher number is associated to the circle closest to the best condition. In this way simplicity of use for the listener is preserved and, without realizing it, he gives a numerical score to each sound system attribute.

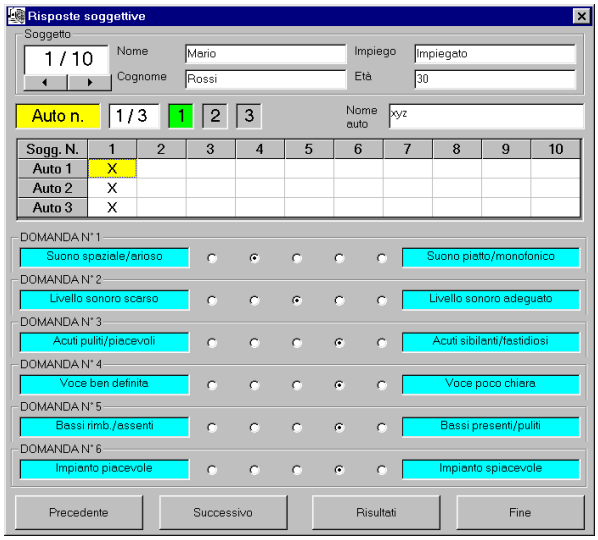

Fig. 2

PRELIMINARY TEST

We also need to establish a method, also statistical, for rating the quality, or reliability, of each individual listener: this has the goal to limit the number of listeners we need to use and to select only listeners with some basic preparation and predisposition for listening to music. In this case, we implemented a very simple, but discriminating, procedure that was capable of providing immediately useful information about each listener, classifying them with respect to the others: it is possible to immediately see how good they are, by the scores they have given during the test. Having constructed an "a priori" statistic that defines the acceptability limit for each listener, one can immediately establish whether a specific listener is to be accepted or not. The a priori statistic is nothing else but the data relative to a consistent number of preliminary tests made by 40 listeners: this gave rise to a distribution (fig. 3) that is not perfectly Gaussian: on the contrary, it presents three modes. This means that there is a small group of high quality listeners, a second more numerous group of medium quality listeners and an average group of bad listeners who necessarily have to be excluded.

The average score is 23.5. To select only the best listeners, the maximum acceptable score was fixed at 20. As you can see in fig. 5, which shows the individual scores, only 13 out of 40 subjects were below this threshold, and therefore only these should be used for measuring the IPA. For the admission of other listeners, it is sufficient to assess their reliability with this preliminary test, and then compare their score: only if this is less than or equal to 20 the listener should be admitted to measuring the IPA.

Preliminary Test Procedure

We can state beforehand that such a preliminary test can be performed on paper, without the use of computer, or, better, based on a computer interactive questionnaire: in this latter case, a software program could be set up that would acquire the responses of the listener under examination and, at the same time, using a sound card, plays the sound reproduction system, amplifier and acoustic boxes, a set of sound samples with artificial alterations.

The preliminary test is very simple, but powerful: the listeners must respond appropriately to a series of questions, demonstrating that they are capable of recognizing the artificial effects added to the musical signal.

In particular, they are presented with the following sound samples:

Sample 1	Original sample, not filtered
Sample 2	Mixed signal from stereo to mono
Sample 3	Low-pass filtering, 6dB/oct to 2000 Hz
Sample 4	High-pass filtering, 10dB/oct a 500 Hz
Sample 5	Distortion (4% THD)
Sample 6	Copy of SAMPLE 1 for a consistency test

For each sample presented, they are asked to respond to the following questions:

Question 1	Distorted	Not Distorted
Question 2	Treble Boosted	Treble Reduced
Question 3	Bass Boosted	Bass Reduced
Question 4	Stereophonic	Monophonic

The matrix of responses provided by an ideal listener is as follows:

Sample	1	2	3	4	5	6
Question 1	9	9	9	9	1	9
Question 2	5	5	9	5	5	5
Question 3	5	5	5	9	5	5
Question 4	1	9	1	1	1	9

The global score for each subject is obtained by summing the deviation of each response from its ideal value: this gives rise to the distributions in (fig. 3, 4). Relative to each question, the choice of five options is offered, associated to a graduated scale from 1 to 9, between an attribute and its complement, or contrary, such as "distorted" and its complement "not distorted." For sample 1, the "perfect" listener must choose "not distorted" with a value of 9: the deviation between the listener's preference and the ideal value, as shown in (tab. 3), is 0. So, the ideal listener must assign 0 as the global score.

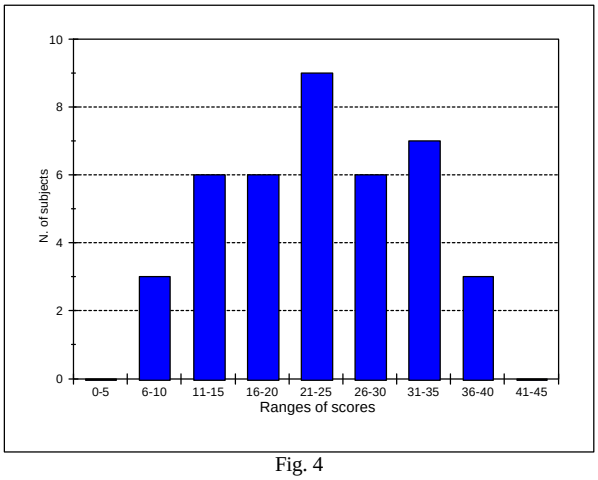

Fig. 4

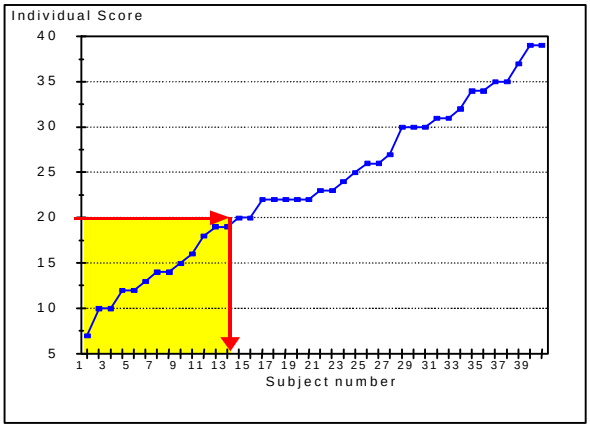

Fig. 5

IPA: CALCULATION ALGORITHM

6 subjective attributes are explored, making use of the following questions (see fig. 2 for Italian equivalence):

N.	Left Attribute	Right Attribute
1	Spatial Sound	Flat Sound
2	Insufficient Level	Sensible Level
3	Clean Treble	Sibilant Treble
4	Clear Voice	Unclear Voice
5	Weak or overshoot bass	Clean and powerful bass
6	Pleasant sound system	Unpleasant sound system

First of all, a fuzzy logic is employed for converting the discrete responses to each of the 6 questions (see fig. 2) to a score x_i, ranging between 1 to 9, associated with the i-th subjective attribute. The fuzzy logic is simply implemented defining a matrix of scores, which give the score values for each of the 5 possible responses to each of the 6 questions. All the questions are monolateral, meaning that the maximum score is on one side (for example, on the left for questions # 1,3,4,6 and on the right for questions # 2 and 5): the maximum score is always 9, and the minimum is 1, with step 2.

The method can easily be adapted also to "centered" questions (that is, attributes specifying opposite defects, such as "too much bass" and "too little bass"), in which case the score is maximum in the middle and minimum at the sides. In this work, however, only monolateral questions were employed, and thus the fuzzy logic could have been substituted by a simpler linear mapping of the circled responses to a discrete numerical scale from 1 to 9.

The scores are subsequently weighted, by multiplication with proper weights: w_1, w_2, w_3, w_4, w_5, w_6, are the relative weights of each attribute. It must be that:

$$\sum_{i=1}^{6} w_i = 1 \qquad 1)$$

While the IPA, provisionally called "static", relative to the j-th listener becomes:

$$IPA_j^{static} = \sum_{i=1}^{6} x_{i,j} \cdot w_i \qquad 2)$$

where the value $x_{i,j}$ is the score corresponding to the choice made, and is included between 1 and 9, relative to the i-th attribute from the j-th listener. The total static IPA due to the contribution of all the listeners, whose number we identify by the letter K, becomes

$$IPA^{static} = \sum_{j=1}^{K} \frac{IPA_j^{static}}{K} \qquad 3)$$

As regards the number K, there is no limit, in the sense that the more listeners contributing to the statistic, the more reliable it is: experimentally, it has been revealed that the minimum number, and thus optimal, to give credibility to the statistics is **K=12**, if these listeners have passed the subjective admission test (preliminary test) with a sufficient score, while it rises to K=30 for randomly chosen listeners who have not taken such a test.

To optimize the calculation of the IPA, we introduced a dynamic correction that, in fact, introduces "dynamic" weights relative to each individual based on their coherence in giving their preferences. Comparing the average score calculated with 2) on the basis of only the first 5 questions, with the preference given to the sixth question, we obtain another weight function, this time belonging to each individual.

So, let S_1, ..., S_K, be the weights relative to each listener. In principle, the following equality must always be valid

$$\sum_{j=1}^{K} S_j = 1 \qquad 4)$$

This implies normalization by dividing for the sum of the weights.

We first define the unnormalized weights T_j as

$$T_j = 9 - \left| x_{6,j} - \frac{\sum_{i=1}^{5} x_{i,j} \cdot w_i}{\sum_{i=1}^{5} w_i} \right| \qquad 5)$$

And thereafter we obtain the normalized coefficients S_j :

$$S_j = \frac{T_j}{\sum_{s=1}^{K} T_s} \qquad 6)$$

Substantially it consists in calculating the "distance" between the partial, static IPA, which is relative only to the first 5 attributes, and the value chosen by the listener relative to the sixth attribute: the greater is this distance, the less weight will be given to that listener, automatically reserving greater weight for the others. In the ideal case in which: a) all the listeners present the same "distance" or b) all the listeners have null "distance", then weights S_i are all equal.

Finally, the "dynamic", and thus real IPA, is calculated as follows

$$IPA = \sum_{j=1}^{K} \sum_{i=1}^{6} x_{i,j} \cdot w_i \cdot S_j \qquad 7)$$

This number is in the SAE scale and provides a grade between 1 and 9. It is possible to know the score relative to each subjective attribute, always on the usual SAE scale from 1 to 9, due to all the listeners

$$P_i = \sum_{j=1}^{K} x_{i,j} \cdot S_j \qquad i = 1,....,6 \qquad 8)$$

When we compare several cars (or several sound systems in a single car), we denote L as the total number of different vehicles (or systems) and the subscript m identifies the m-th vehicle. Now we indicate IPA_m as the IPA calculated for the m-th vehicle, and the same for the score $P_{i,m}$ relative to the i-th attribute and the m-th vehicle.

Calibration of weights w_i

In the preceding sections we mentioned weights w_1, w_2, w_3, w_4, w_5, w_6, relative to each attribute. Now, the value to be given them, with respect to 1), can be derived from past experience [3,6], which is to say be going over the historical archive, mediating it through acquired knowledge and the unavoidable laws of physics.

Guided by such implications, we assigned numerical values as follows:

$w_1 = 0.2$ $w_2 = 0.1$ $w_3 = 0.1$
$w_4 = 0.15$ $w_5 = 0.25$ $w_6 = 0.2$

SOFTWARE IMPLEMENTATION

The methodology described can be profitably employed to evaluate the audio system in the car, as already introduced, by providing a paper form to fill out during the listening session and handling the collected data with the special software described here.

To automate the method for collecting and handling the data, so as to achieve a presentation of the rapidly interpreted results, we use a processing software program capable of providing the IPA in interactive mode with scores for each attribute.

First we constructed the format of the forms to be filled out (fig. 1, 2) and from there we passed on to a software implementation, including all the parts of recursive calculations, as well as the graphic representation of the results (fig. 6).

For greater user support, the program is provided with on-line "Help".

Program Structure

The software program was written in Visual Basic 6.0.

We built a setup file that automates installation, once launched on the computer. This setup can be supplied on 2 floppy disks.

The program manages the user interface, displaying an initial setup form (fig. 1) and one for questions, similar to the paper form (fig. 2) on the screen. At the same time, it stores the data entered and processes it following the algorithm described above, giving visible feedback of the completion of the tasks through a matrix of X-signs, which show how many questionnaires were already completed. Let's see how the program executes the various operations in detail.

Initial Setup Form

The first screen (fig. 1) allows setting the number of subjects and the number of vehicles (default values are 10 subjects and 3 vehicles).

You can also choose, in the case that you are continuing to process a previous session, to overwrite the old data or to add the new data following the old. In reality, it's possible to modify the pre-existing data in either case. The difference is given only by the number of subjects, which in the second case is added to those already present, while in the first case, it is assumed to be equal to the larger of the number already present and the number indicated in the program start-up screen.

You then must select the name of the file. If you choose the name of an existing file, you will receive a warning before it is opened.

By pressing "OK" you pass to the screen for entering the subjective responses.

Response Table Form

A rectangular input screen (fig. 2), similar to the paper form filled in by the various listeners, shows the entry status of the total data table through the appearance of crosses corresponding to the group of responses from each subject for each vehicle. If a cross does not appear when moving to the next subject/question, this indicates that some data entry has been skipped.

It is also possible to select a subject/car pair simply by "clicking" with the mouse on the box corresponding to the completion screen.

Only when the screen is completely filled with crosses is data entry complete.

It is a good idea to diligently finish data entry, entering the first and last names for each subject and the name of type of each vehicle tested.

At this point, you can press the "Results" button that displays the computed values of IPA in graphics and tabular formats.

Once you exit from the graphics, you can exit from the program: all the responses and results are saved in the results file (ASCII, easily imported in spreadsheets), from which they can be reloaded in a later session. In this way, it is possible to begin the analysis with a small number of subjects and cars and add additional data at a later time.

Graphic Display of the Results

The score obtained from each question, and the total score, are shown for each car. You will recall that the score for each single question is being weighted for giving the total score, as we discussed above. The weights are dependent on the question (static weights) and on the particular response of each subject to each car. This means that the weights are not influenced by the answers given for the other cars: as a consequence, given a certain set of subjective responses, each car always receives the same total score, independently of the score of the other cars. The total score is always between 1 and 9.

The graphics could be saved in the "clipboard," using the copy function, which also loads the data table. In this way, both the data table and the graphic can be imported by Excel (using the "paste special" function).

The graphic can also be printed directly.

The IPA relative to each car, called "Tot. Score" appears at the foot of the graphic of the results, while the score relative to each attribute and each car is associated to each bar of the histogram.

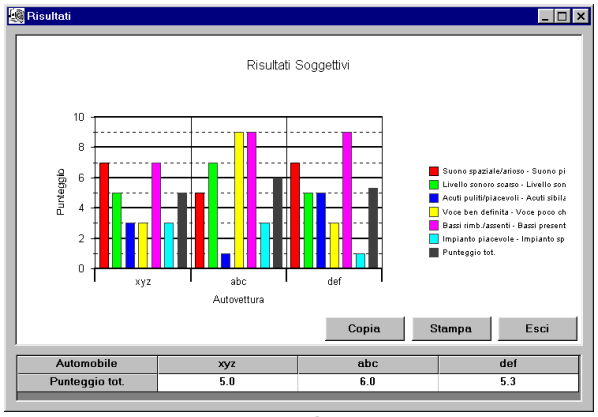

Fig. 3

Selection of the Musical Pieces

In order to arrive at a measurement of the IPA, you have to provide the listeners with a certain number of musical pieces that have to be reproduced through the sound systems being evaluated.

In fact, it is very important that all the questionnaires are compiled after listening to the same music pieces.

So we need to standardize the choice of music pieces. Also in this case, there is no precise rule, but only logical implications and those dictated by experience.

First of all, it's necessary to keep in mind that the evaluation has to be linked to reality, so the musical selections should be selected from those available at the time, with excellent recording quality. A good rule is to set up a series of homogeneous selections taken from various musical genres: rock, classical, jazz, rhythm & blues and pop. In addition, the time associated with the evaluation must be taken into account: it has been established that a maximum playback time of 20 minutes for each car is optimal. This allows the listener to adapt to the environment, to take note of any differences and, thus, make up their mind about their preferences. With this total time period available, we deduce that the number of musical selections, assuming an average length of 80 seconds each, is 15. These selections, digitally extracted from the master, are burned onto a CD, which is reproduced on the player in the system to be evaluated.

CONCLUSIONS

This paper presents a simple method for transforming the results of subjective listening tests in a single weighted score, conforming to the SAE scale (1 to 9). The results are thus dimensionally compatible with other similar scores employed in the automotive industry for rating the subjective appreciation of other human perceptions (such as noise, seat comfort, drivability,, etc.).

A double-weighting method was developed: the first set of weightings are associated with the relative importance of the 6 subjective attributes being evaluated. The second set of weightings are instead dependent on the coherence expressed by each subject

when judging each car sound system. In a previous temptative version of this method, the weighting was the same for all the cars judged by the same subject, but it resulted that often the subjects get bored after doing too many listening tests, and thus their reliability is not constant, but tends to decrease. In practice, the method revealed to be very robust, even when the subject panel is not selected by means of the preliminary discriminatory test.

The reliability of the method was verified by repeating the comparative subjective assessment of three sound systems (three cars of segment B) making use of two limited-size, independent groups of subjects: 8 selected subjects chosen among technicians and sound designers working at ASK Automotive Industries (panel #1), and 23 students of the Engineering Faculty of the University of Parma (panel #2). The first panel passed the preliminary discriminatory test, while the second panel was left absolutely unselected, leaving to the IPA method the task to weight-out unreliable subjects.

The following table shows the IPA values obtained from these two panels of subjects:

Sound system n.	Panel #1 (experts)	Panel #2 (students)
1	3.1	3.9
2	5.6	5.7
3	7.6	6.9

It can be seen that the values are not perfectly coincident, nevertheless the three sound systems were ranked in the same order with both panels of subjects. It must be said that these three sound systems were very different, they were easily identified also by untrained listeners.

Regarding the difference among the responses to the single questions, fig. 6 show a comparison between the graphics results obtained in the two cases.

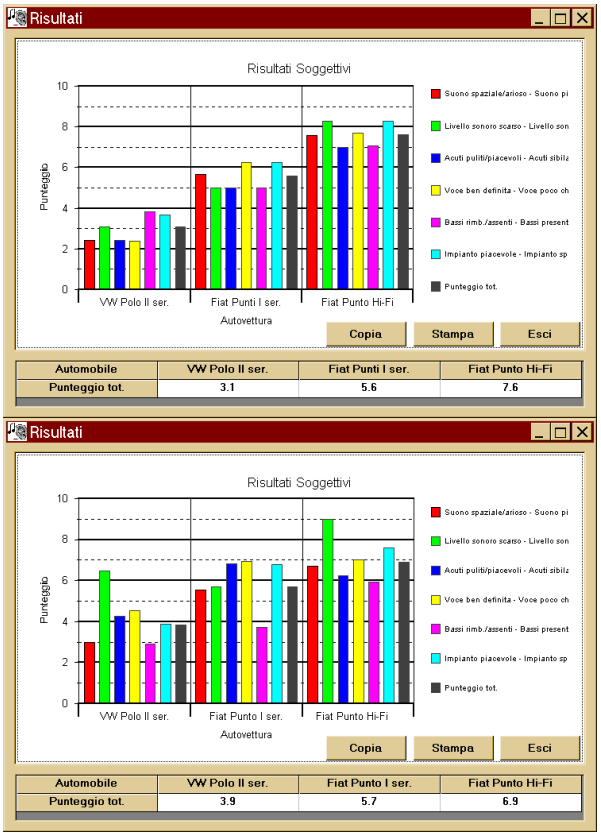

Fig. 6 – Listener's Panel #1 (above) and Panel #2 (below)

From these comparative results, it can be concluded that the IPA method is reasonably robust, and is capable of giving stable values of the global score also if some unreliable subjects are included in the panel of listeners, although of course, like any other evaluation method based on subjective responses, the use of a selected panel of trained listeners is always advisable.

ACKNOWLEGDMENTS

This work was supported through a research convention between ASK Industries, Reggio Emilia, Italy and the University of Parma, co-funded by the Italian Ministry for University and Research (MURST) under the grant MURST-98 #9809323883-007.

References

[1] A. Farina, E. Ugolotti, "Automatic Measurement System For Car Audio Application", Pre-prints of the 104rd AES Convention, Amsterdam, 15 - 20 May, 1998.

[2] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique", ", Preprint of the AES 108th Convention, Paris, February 2000.

[3] A. Farina, E. Ugolotti, "Subjective comparison of different car audio systems by the auralization technique", Pre-prints of the 103rd AES Convention, New York, 26-29 September 1997.

[4] E. Granier, M. Kleiner et al., 'Experimental auralization of car audio installations' Journal of the Audio Engineering Society, vol. 44, n. 10, 1996 October, pp.835-849.

[5] A. Farina, E. Ugolotti, "Subjective comparison between Stereo Dipole and 3D Ambisonics surround systems for automotive applications", 16th AES Conference, Rovaniemi (Finland) 12-14 April 1999.

[6] A. Farina, E. Ugolotti, "Subjective Evaluation Of The Sound Quality In Cars By The Auralisation Technique", Proc. of 4th International Conference and Exhibition "Comfort in the automotive industry" - Bologna (Italy) October 2-3, 1997.