Audio Engineering Society

Convention Paper 6356

Presented at the 118th Convention
2005 May 28–31 Barcelona, Spain

Comparison of different listening systems for speech intelligibility tests

Andrea Azzali[1], Paolo Bilzi[1], Eraldo Carpanoni[1], and Angelo Farina[1]

[1] *Industrial Eng. Dep., Universitá di Parma, Parco Area delle Scienze-43100 Parma-Italy*

Correspondence should be addressed to Azzali (azzaliandrea@libero.it)

ABSTRACT

Being able to have a fast and reliable evaluation of speech intelligibility inside cars is of outmost importance during the interiors design phase. Performing subjective tests inside car compartments is a time consuming task and therefore not suitable for industrial processes. Moreover, comparison of different car fittings and database implementation cannot be performed in an easy and fast way. The effectiveness of virtual listening systems was therefore investigated in the context of a research project that involves both UNIPR and an important carmaker. A preliminary subjective evaluation session in real cars was carried out. Then two different virtual listening systems were compared to investigate the best configuration for the intelligibility tests. In this paper a comparison of the systems is presented and the provided experimental results show that the tests performed in the listening room are consistent with the ones carried out in car compartments.

1. INTRODUCTION

Facility of communication between passengers inside car compartments is becoming an increasingly relevant part of customers global comfort and satisfaction. Furthermore, a good intelligibility is required for security reasons: the driver should be able to listen to and talk to the other passengers without turning his head. Finally, the new infotainment systems require good intelligibility in car compartments. Therefore, optimal listening conditions are of paramount importance for carmakers.

One of carmakers needs is the ability to evaluate and compare the speech intelligibility inside different car compartments in order to design the best cost/performance interior solution. This allows both customers needs satisfaction and project and production costs optimization. Therefore, the possibility to compare different car fittings in a fast, reliable and repeatable way is fundamental for carmakers.

Performing subjective tests inside car compartments is a time consuming task and not suitable for industrial processes. In order to reduce the time necessary to evaluate speech intelligibility and to provide the designers with a comparison tool, objective parame-

ters are usually exploited. The most widely used parameter for speech intelligibility evaluation is the Articulation Index [1], but in the last years other objective parameters, such as the Speech Transmission index [2], derived from architectural acoustic, are used [3]. An alternative solution to fast evaluate speech intelligibility is to perform subjective tests inside virtual environments: in a few minutes it is possible to ask people to evaluate different car settings.

This paper describes the very first steps of a wider research project that involves both UNIPR and an important carmaker. One of the goals was to compare and validate different virtual listening environments for intelligibility reproduction. Two different virtual listening systems were evaluated: a special listening room and headphones plus subwoofer. Subjective tests were performed both in the virtual environments and in car compartments in order to validate the effectiveness of the reproduction systems. The obtained results demonstrate that the relative performance of the different cars is the same in the real conditions and in the listening room.

A similar work [4], regarding room size and source distance perception, was carried out in collaboration with the University of Sidney and presents results that are coherent to the ones obtained in the context of this research.

2. SUBJECTIVE TESTS CONSTRUCTION

The first session of tests was performed inside three different cars (in the following labeled A, B, C) and three different situations were considered: engine off , 90 km/h speed and 110 km/h speed. The tests were built on the phonetically balanced list of 200 Italian words firstly published in 1993 by Turrini et al. [5] and the response method was open. An artificial mouth simulator, with a directivity similar to the human directivity [6] and mounted on a simplified head and torso model, was used as the speaker. During the engine off tests, both nonsense and meaningful words were used; while only meaningful words were used in the engine on tests. Each word was preceded and followed by a sentence: the carrier sentence. Each listener had to listen to and then recognize 10 meaningful words and 15 nonsense words in three different car compartments and in the three different speed conditions previously mentioned. The percentage of correctly recognized words was the main investigated parameter. Moreover, people were asked to rate the difficulty level of each listened word comprehension.

In order to evaluate the performance of the virtual reproduction systems, the same test proposed in the three cars was carried out into two virtual listening systems. The list of words and the background noise at 90 and 110km/h were recorded inside the three cars to obtain a database of useful tracks to be exploited during the tests inside the virtual environments. The *B&K 4100* dummy head was used for the binaural recording.

3. THE VIRTUAL LISTENING SYSTEMS

The main problem during the set up of a listening test is to understand which is the most useful and performing reproduction system for that kind of test. Headphones are currently the most widely used tool for listening tests proposal and comparison. In acoustics literature, several works on the real capability of headphones to correctly reproduce the sound field were presented. In this work, headphones and a listening room of novel conception were compared and their effectiveness in properly reproducing intelligibility inside cars was investigated. One of the main addressed questions was: how well can the two-channel recording technique and the different auralization systems based on that one, maintain hearing features?

Before presenting the construction of the virtual listening room, a brief description of the employed binaural reproduction techniques will be provided.

3.1. Headphones

Stereo headphones are the most intuitive tool to reproduce binaural recording. They should recreate at the ears a pressure equal to the recorded one, maintaining the separation between the two hearing channels (in this case no "cross talk" paths are present) and without being affected by the room response. In order to achieve this goal, the transfer function from the headphones signal to the inner ear is usually measured using the same dummy head used during recording. Then, two inverse filters are calculated and applied to the binaural signal to flatten the frequency response between the headphones and the microphones inside the dummy head, and

to exactly reproduce the signal recorded at the inner ear. Reproduction would be very realistic if the head used for recording (or IR measurement) was the same as the listener head. For obvious reasons it is necessary to use a standard dummy head for recording, and this affects the reproduction in a non negligible way. Moreover, the fact that the reproduced sound image is rigidly bounded to eventual little movements of the listener head and the fact of wearing an object on the head,which shields the listener from the natural extern background noise, represent psychoacoustic negative artifacts. For the tests *Sennheiser HD580* headphones were used.

3.2. The listening room

The listening room is a reproduction system developed to obtain better results from the two channel recording technique. Usually, carmakers perform binaural recording rather than multi-channel recording (i.e. soundfield microphone technique) as this is a less time-consuming approach. Making the listening room able to accurately reproduce the binaural recording without loss of perceived frequency response, spatiality, and sound or noise localization is a hard task. A specific loudspeaker configuration is proposed in this paper.

The double stereo-dipole loudspeaker configuration based on the cross talk cancelation algorithm, gives better results than the standard stereo system in the reproduction of stereophonic recording. In particular, great performance of stereo dipole can be achieved with recording made inside large spaces and with commercial tracks. Several problems were encountered instead in the stereophonic reproduction of a recording made inside the car cockpit because of several reasons. Mainly, the high reflective surfaces (lateral and frontal windows) and the very small size of the cockpit, make strongly correlated to each other the two signals recorded from a dummy head. This produces a loss of capability in localizing sound and noise sources. The recording inside a car has also a great frontal fraction that a double stereo dipole system is not able to accurately reproduce for this special kind of application. So a new loudspeaker displacement, starting from a double stereo-dipole system, was developed. Two additional loudspeakers were positioned in lateral position for stereo enhancing and two more in front of the listener for room size control.

In the following the technique used for the construction will be explained. The results show that the system reproduces a car cockpit sound field closed to the real one.

3.2.1. Cross Talk cancelation

This technique uses a two by two matrix of four filters, calculated so that the system cancels the contribution of the left speakers to the right ear and viceversa. This matrix H is obtained by inverting the original matrix C of the speakers to ears transfer function previously measured. Kirkeby et al.(1998) [7] found that a configuration with a 10 degree interval between loudspeakers as seen by the listener, minimizes the ringing artifacts in the cross-talk cancelation filters, and expands the area in which the cross-talk cancelation is effective (allowing greater listener head movements). This method gives of course a more natural sensation, not relying on a strange and close source like a headphones pair. Moreover, the portion of sound coming from the front (in most recordings it coincides with the direct field of a source) is spatially correctly reproduced, not only in the neighborhood of the ears, but on a wider area, inducing a natural spaciousness sensation when slightly moving the head. Figure 1 shows the cross-talk phenomenon in the reproduction space.

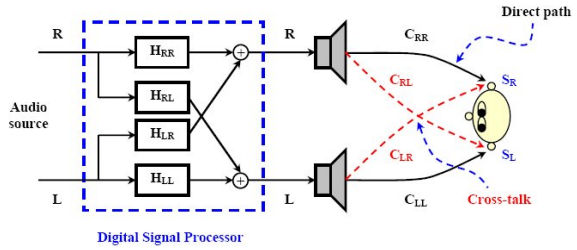

Fig. 1: Scheme of stereo dipole

3.2.2. Double stereo dipole

The double stereo-dipole adds a rear pair of loudspeakers to the normal stereo-dipole configuration. The processing is made by two dipole matrixes, *H-front* and *H-rear*, calculated by inverting independently the two direct matrixes *C-front* and *C-rear*. With this approach, the ear pressure induced by the two stereo-dipoles, front and rear, well approxi-

mates the one in ideal conditions (listener head coincident with measurement head, listener head perfectly positioned and still, ideal reproduction environment). Moreover, the double stereo-dipole gives also for sound coming from behind, the same advantages previously described, which the single dipole gives only for sound coming frontally: hence it is supposed to be more realistic in situations in which rear sound is particularly important.

3.2.3. Kirkeby inversion of a cross-talk stereo system

The 4 cross-talk canceling filters f, which are convolved with the original binaural track, have to be designed so that the signals collected at the ears of the listener are identical to the original signals. Imposing that $p_l = x_l$ and $p_r = x_r$, a 4x4 linear equations system is obtained. Its solution yields:

$$f_{ll} = (h_{rr}) \bigotimes InvDen \quad (1)$$

$$f_{lr} = (-h_{lr}) \bigotimes InvDen \quad (2)$$

$$f_{rl} = (-h_{rl}) \bigotimes InvDen \quad (3)$$

$$f_{rr} = (h_{ll}) \bigotimes InvDen \quad (4)$$

$$InvDen = InvFilter(h_{ll} \bigotimes h_{rr} - h_{lr} \bigotimes h_{rl}) \quad (5)$$

The problem is the computation of the $InvFilter$ (denominator), as its argument is generally a mixed-phase function. In the past, the authors attempted [8] to perform such an inversion employing the approximate methods suggested by Neely&Allen [9] and Mourjopoulos [10], but now the Kirkeby-Nelson frequency-domain regularization method is preferentially employed, due to its speed and robustness. A further adaptation of the previously published work [11] consists in the adoption of a frequency-dependent regularization parameter. In practice, the denominator is directly computed in the frequency domain, where the convolutions are simply multiplications, with the following formula:

$$C(\omega) = FFT(h_{ll}) \times FFT(h_{rr}) - FFT(h_{lr}) \times FFT(h_{rl}) \quad (6)$$

Then, its complex inverse is taken, adding a small, frequency-dependent regularization parameter:

$$InvDen(\omega) = (Conj[C(\omega)]) \div (Conj[C(\omega)] \times C(\omega) + \varepsilon(\omega)) \quad (7)$$

In practice, $\varepsilon(\omega)$ is chosen with a constant, small value in the useful frequency range of the loudspeakers employed for reproduction (80 - 16k Hz in this case), and a much larger value outside the useful range. A smooth, logarithmic transition between the two values is interpolated over a transition band of 1/3 octave.

Fig.2 shows the user interface of the software developed for computing the cross-talk canceling filters.

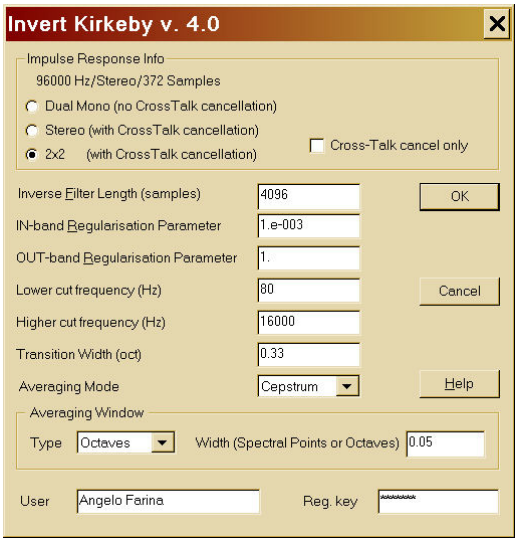

Fig. 2: User interface of the inverse filter module

This software tool was implemented as a plug-in for *Adobe Audition*, and it can directly process a stereo impulse response (assuming a symmetrical setup, so that $h_{ll} = h_{rr}$ and $h_{lr} = h_{rl}$), or a complete 2x2 impulse responses set, obtained by placing the binaural IR coming from the left loudspeaker, followed in time by the binaural IR coming from the right loudspeaker. In both cases, the output inverse filters are in the same format as the input IRs. The computation is so fast (less than 100 ms) that the optimal values for the regularization parameters can be easily found by employing a trial-and-error method.

4. SPEAKERS POSITIONING

As a further trick for resonance lowering , the axis of

symmetry of the loudspeakers array was not aligned with the room, nor was the listener positioned in the room center. According to the double stereo-dipole layout, two *Quested Studio Monitor Model WH2108* loudspeakers were placed at 1.82 m in front of the listener position and two *Quested Studio Monitor Model F11P* loudspeakers in stereo-dipole configuration were positioned at a distance of 3.14 m behind the listener. Moreover, in this new configuration some hearing and emotional aspects were enhanced by using additional loudspeakers. Firstly, the spatiality perceived during reproduction results partially corrupted by the critical conditions presents during recording inside the car. To overcome this problem a pair of *JBL TLX 181* loudspeakers were placed in lateral position at 1.13 m. A great enhancing of the spatiality was achieved. Finally, two self made loudspeakers were introduced in front of the listener at 1 m height in order to optimize the acoustic directivity. An emotional aspect was also corrected in this way.

Typically, the reproduction in the listening room gives a perception of larger environment dimensions than in cars. In the cockpit the sound is perceived as close to the body and emotions are strong during the hearing. Introducing a pair of loudspeakers so close to the body, the energy ratio and the delay between the direct sound and the reflected path are made closer to the real condition. The system uses a *Quested VS Series Active* subwoofer for the low frequencies.

5. FREQUENCY AND TIME EQUALIZATION: INVERSE FILTERING AND TIME ALIGNMENT

Typically, the reproduction system does not have a true flat response and distorts the harmonic, spatial and dynamic behavior of the recorded tracks. Indeed, loudspeakers and environment responses affect the hearing. Therefore, a set of inverse filters for each system were calculated using the *Aurora* plug-in hosted by *Adobe Audition*. The impulse responses between each loudspeaker and listening seat were calculated using sweep stimulating signals. The inversion was performed using the previously mentioned Kirkeby algorithm, implemented as a plug-in of *Aurora*. The cross talk cancelation algorithm and the time alignment and inverse equalization filters are implemented on a PC by using *Audiomulch* software, a 10 output channels *Event Layla* audio card and the *Voxengo Pristine Space* convolver.

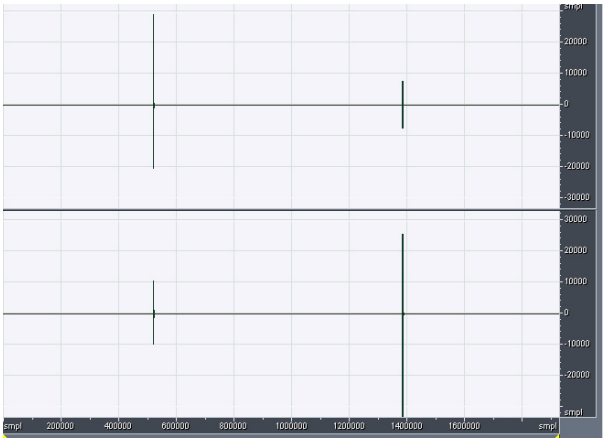

Fig. 3: Cross Talk cancellation measured in the listening room

The listening room setting procedure is based on different binaural impulse response measurements made by using a *B&K* dummy head and torso placed on a listener position and *Audition* software with *Aurora* plug in. A series of binaural impulse response measurements was done to obtain the inverse filter for flattening the response of the other stereo loudspeaker and subwoofer.

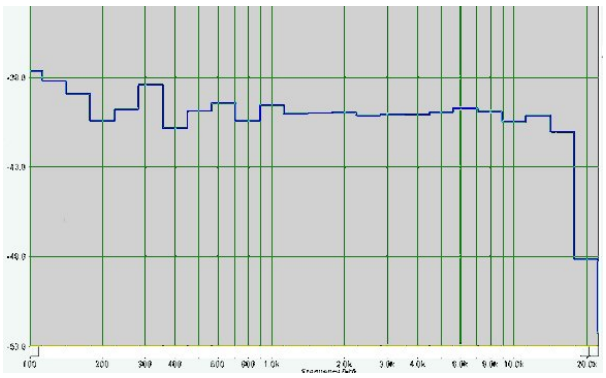

Fig. 4: Frequency response at the listener position

Finally, the time alignment and sound pressure level calibration were performed for the loudspeaker when

the digital filters were working. The frontal stereo-dipole system was set to have 3 dB more than the stereo system in order to obtain a good full system response. In fig.3 , the impulse responses of the complete system for a stereo signal (the left and right channel are stimulated separately) are presented to show the cross talk cancelation effect.

The frequency response at the listener position (Fig.4) are measured and a noise sample track was recorded in order to compare it to the original track (Fig.5).

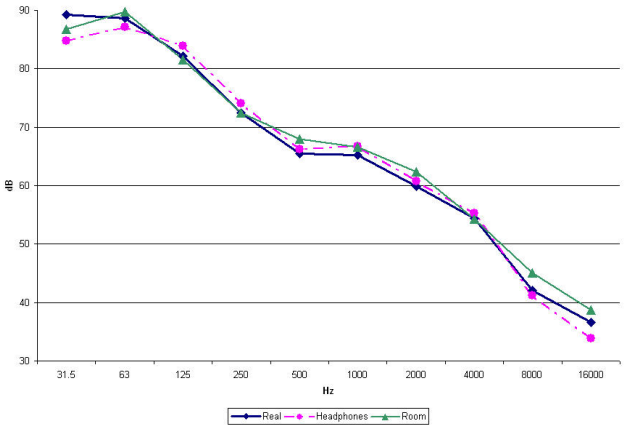

Fig. 5: Comparison between real noise and virtual noise spectrum in listening room and headphones

6. TESTS RESULTS

For every car and for every speed condition, the words scores and the mean opinion scores of the comprehension difficulty level were statistically analyzed. Fig.6 shows the number of correctly recognized words.

In the figure it can be seen that in the engine on configurations, the results obtained inside the virtual environments are consistent with the real case. When the engine is off there is a gap between the words scores of the virtual systems and the ones of the real situation, but the relative scores between the cars are unchanged. Indeed, adding an offset to the words scores of the virtual systems, as shown in Fig.7, gives to the engine off case results a good correlation with the real situation. The offset for each virtual environment was computed in order to

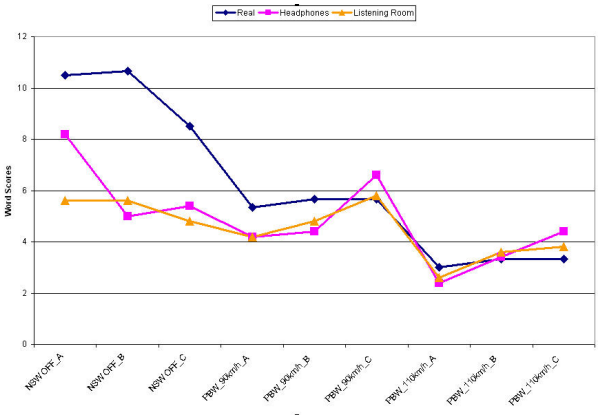

Fig. 6: Word scores in real case and in virtual environment

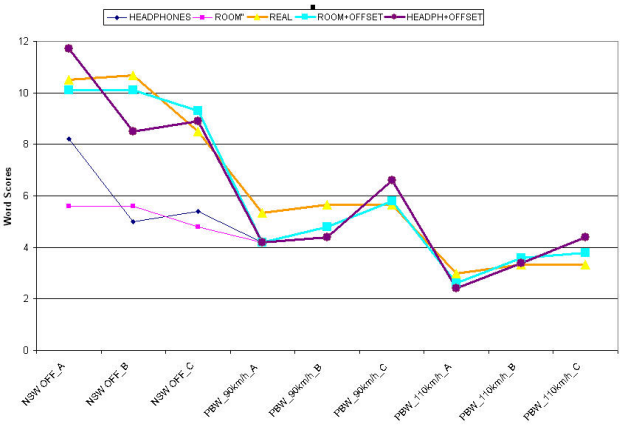

Fig. 7: Word scores in real case and in virtual environment with offset

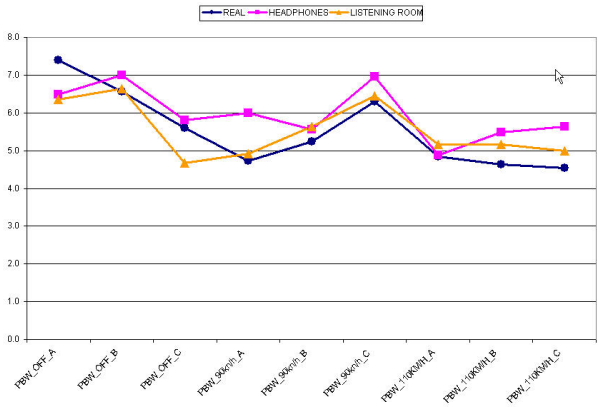

Fig. 8: Word scores in real case and in virtual environment

minimize the gap between the real and the virtual configuration. It can be easily seen that, while in the engine on case the two virtual systems are consistent with the real case, when the engine is off, the listening room provides better results. Moreover, the mean opinion scores of the comprehension difficulty level (fig.6) gave a good correlation between the real case and the virtual systems.

7. CONCLUSIONS

A comparison between two different virtual listening system for intelligibility tests inside car compartments was carried out. The two different systems were a headphones plus subwoofer and a special listening room. The special listening room was designed to simulate the car cockpit environment. The same listening room was also used in a research project which goal was to develop an objective index to predict the perceived quality of car stereos [12]. During the tests both the virtual systems showed a good correlation with the real system, but in the engine off test case, the listening room provided better results. Probably this is due to the fact that intelligibility is mainly affected by the background noise and both the virtual systems can reproduce it with high fidelity. When the engine is off instead, intelligibility is also affected by sound spatiality and that effect is better reproduced by the listening room rather than by the headphones. The listening room effectiveness as a playback system is better than the headphones when the complexity of the perceptive aspects that should be recreated becomes bigger. That was also shown in a research project that involved Parma and Sydney Universities in which it was shown that it is possible to recreate the room size perception and source distant perception by using a listening room, but not by using headphones [4].

8. ACKNOWLEDGEMENTS

The authors want to express their gratitude to ASK industries (Reggio Emilia -Italy-) for the fundamental support given with the access to their special listening facilities.

9. REFERENCES

[1] ANSI S3.5-1969, *Method for the Calculation of the Articulation Index*

[2] IEC 60268-16:2003, *Objective Rating of Speech Intelligibility by Speech Transmission Index*

[3] A. Farina, F. Bozzoli, *Measurement of the speech intelligibility inside cars*, Pre-prints of the 113th AES convention, Los Angeles, 5-8 october 2002

[4] A. Azzali, D. Cabrera, A. Capra, A. Farina, P. Martignon, *Reproduction of auditorium spatial impression with binaural and stereophonics sound systems*,Pre-prints of the 118th AES convention, Barcelona, 28-31 may 2005

[5] Turrini M., Cutugno F., Maturi P., Prosser S., Albano Leoni F., Arslan E., *Nuove parole bisillabiche per audiometria vocale in lingua italiana*, Acta ORL Italica, 13, 1993, pp.63-77

[6] F.Bozzoli, A. Farina, *Directivity balloons of real and artificial mouth simulators for measurement of the Speech Transmission Index*, Preprints of the 115th AES convention, New York, 10-13 october 2003

[7] O. Kirkeby, P. A. Nelson, H. Hamada - "The "Stereo Dipole"-A Virtual Source Imaging System Using Two Closely Spaced Loudspeakers" - JAES vol. 46, n. 5, 1998 May, pp. 387-395

[8] A. Farina, F. Righini, *Software implementation of an MLS analyzer, with tools for convolution, auralization and inverse filtering*, Pre-prints of the 103rd AES convention, New York, 26-29 September 1997

[9] S.T. Neely, J.B. Allen, *Invertibility of a room impulse response*, J.A.S.A, (1979) no. 66, 165–169

[10] J.N. Mourjopoulos, *Digital equalization of room acoustics*, JAES, (1994) no. 11, 884–990

[11] A. Farina, E. Ugolotti, *Spatial equalization of sound systems in cars*, Proc. of 15th AES Conference, Copenaghen, 1998

[12] A. Azzali, A. Farina, G. Rovai, G. Boreanaz, G. Irato, *Construction of a car stereo audio quality index*, Pe-prints of the 117th AES convention, San Fransisco, 28-31 october 2004