A performance comparison of Home Usage Testing and Central Location Testing in small impairment listening tests.

Søren V. Legarth [1], Jesper Ramsgaard [1], Guillaume Le Ray [1], Nick Zacharov [1]

[1] SenseLab, DELTA, Denmark.

svl@delta.dk

Abstract

The performance of listening tests can be a laborious process and requires significant preparation and control to ensure good data quality . This is particularly true for small impairment listening tests performed in accordance to recommendation ITU-R BS.1116-1 [1] in dedicated listening rooms under calibrated conditions. The usage of such central location testing (CLT) methods is time consuming, costly and limited to the physical location. An alternative type of test is the home usage testing (HUT) whereby the assessors evaluate a product in normal usage at home, as commonly encountered in consumer evaluations. This paper provides a comparative analysis of the two methods in a small impairment study of audio codec performance.

Index Terms: Listening tests, audio coding, ITU-R BS.1116-1, small impairment, home usage testing, central location testing.

1. Introduction

The audio industry often requires the use of listening tests to evaluate the quality of sound reproduction equipment and new technologies, such as audio coding, in order to establish the perceptual audio quality , as extensively discussed in [2]. Different recommendations have been published by the International Telecommunication Union (ITU), for example, to facilitate and standardise these tests. Such recommendations provide guidance on how to design and conduct listening tests in order to ensure good data quality . Conducting thorough listening tests is the relatively slow process with high associated cost, which grow rapidly with increasing number of test variables, often rendering these tests impractical. In particular small impairment listening tests as used in the evaluation of codecs, as prescribed in ITU-R BS.1116-1 recommendation [1] are performed in dedicated listening rooms under calibrated conditions. The usage of such so-called central location testing (CLT) methods are time consuming, involved and also limited to the physical location of the lab with associated logistic constraints (e.g. assessments performed by one assessor at a time). For example with a test configuration comprising of 10 sound samples processed via 3 systems (codecs), and evaluated by 20 assessors (normal size test) requires ~50 lab hours.

An alternative approach, often employed in consumer studies is known as home usage testing (HUT), whereby assessors, typically consumers, evaluate a product in normal usage at home. Normally employed for the assessment of consumer goods such as food or personal care products (shampoo, hand cream, etc.), HUT studies aim to be more ecologically motivated compared to the artificial and contrived CLT. A number of studies have been made to compare CLT and HUT results for a range of consumer products, as discussed in [3], [4] and [5]. Depending on the nature and the difference between the CLT and HUT test configuration, so either similar or significantly different results can be gained from two approaches.

The benefits of a HUT approach for consumer product assessment are multiple, including:

- Natural usage environment
- Privacy
- Independence from the lab environment

However, the CLT also provides its own benefits, including:

- High level of experimental control
    - o Stimulus presentation
    - o Data gathering
    - o Security
    - o Focused assessors

Clearly, both approaches have their strong motivations, but also some drawbacks. From the earlier studies reported in [3], [4] and [5], it can not be generalised whether HUT tests yield similar results compared to CLT. This is very much product and case dependent.

The motivations for this study come from a desire to establish whether or not critical assessments of audio quality need to be performed solely in CLT conditions, or whether similar results can be obtained by using more dynamic and flexible HUT approach, in the form of online web based audio quality assessments.

In consumer product testing, however, it is common to evaluate the fully packaged products in a HUT e.g. tooth paste in standard commercial packaging. This compares with a CLT, where only the product itself is tested in a highly controlled and blind testing manner (e.g. well defined quantity of tooth paste on a standard/neutral tooth brush with neutral water and a defined temperature). These differences in testing protocol in themselves can lead to significant difference in results between HUT and CLT.

In the case of audio testing, as we have full control over the stimulus, we are able to control and repeatably control sound

10.21437/PQS.2010-24

reproduction in such a manner that we can present assessors with identical stimuli both in the HUT and CLT situation. This allows us to compare identical stimuli using both testing environments to make a critical assessment and comparison of CLT versus HUT. We can thus test the hypothesis of whether identical audio quality performance evaluations can be obtained in both test conditions.

In order to make the comparison more interesting, we compare an ideal critical central location listening test of small impairments against a pragmatic home usage test setup of identical audio stimuli. The CLT is performed in a standard compliant listening room over calibrated professional/lab grade equipment and loudspeakers. The HUT is performed using the assessors own PC, an arbitrary web browser with moderate grade headphones, USB sound card and subjective calibration.

2. General Methods

The listening tests are based on the ITU-R BS.1116-1 [1] recommendation (a "double-blind triple-stimulus with hidden reference" method). The 12[1] expert assessors, as defined in [6] of the panel chosen for this study have worked previously with a wide range of SenseLab listening tests. and are familiar with related test stimuli and test paradigms. Following the ITU-R BS1116-1 recommendation [1], a full factorial experimental design has been performed. The CLT test has been performed with 2 replicates whereas the HUT has been conducted with only 1 replicate. A double blind random presentation order was provided for each assessor to avoid order effects.

2.1. Sample stimuli

Test material was selected according to ITU-R BS.1116-1 [1], which states that critical samples that stress the codec shall be used. The 10 critical samples are primarily based on known and commercially available samples from EBU Sound Quality Assessment Material recording for subjective tests (SQAM CD) [7], which are generally considered to comprise of critical audio material for codec and similar audio algorithms. The list of selected test material is found in Table 1.

The objects under study were different types of audio codecs, introducing very small degrees of impairment. Due to the nature and degree of the impairment the ITU-R BS.1116-1 [1] test methodology was consider well suited. Additionally , the original PCM test sample was employed as a hidden reference, as control case for assessing assessor performance. Thus in total 3 test systems were used to process each test sample described in Table 1.

The reproduction levels for the stimuli were subjectively adjusted to the same level to ensure direct comparison between tests. The acoustic reproduction level of the test material was originally adjusted by the experimenter to a most comfortable listening level suitable for critical listening without causing discomfort to the assessors. Measurements were performed at listener position (ear height) to document the reproduction level of each reference sound sample in the test. The measurements were performed using calibrated equipment. Measurement results are shown in Table 2.

2.2. SenseLabOnline user interface

For both the CLT and HUT experiments the SenseLabOnline test interface was employed providing tight experimental control of both the stimulus presentation and the gathering of assessor data. Before starting the test, instructions about the test methods and recommendations were provided, which assessor studied prior to the test.

No.	Track	Content	Duration
1		Muted trumpet	13s
2	SQAM_13	Flute	17s
3	SQAM_35	Xylophone	35s
4	SQAM_40a	Harpsicord	23s
5	SQAM_49	Female speech	23s
6	SQAM_50	Male speech	22s
7	SQAM_60	Piano, Schubert	31s
8	SQAM_62	Soprano, Spiritual	31s
9	SQAM_67	Wind Ensemble, Mozart	32s
10	SQAM_69	ABBA	33s

Table 1: *The 10 selected sound samples for the testing.*

Loudspeaker	L_{Aeq} of reference pink noise signal
Left	63.3 dB(A)
Right	63.4 dB(A)

Table 2: *Acoustic measures (A-weighted) for reference pink noise sample measured at listener position at calibrated level.*

The assessors can at will switch between three stimuli: The known reference is always available under the "Reference" button, the item (i.e. a given combination of a programme material and codec) and the hidden reference. The test item and hidden reference are randomly assigned to the two play buttons below the sliders from trial to trial. The assessors are asked to identify the reference and to evaluate the impairments of the item compared to the reference. Any perceived differences between the reference and the item must be interpreted as impairment. The discrete 5-point *basic audio quality* scale was employed for gathering assessor ratings.

2.3. Experimental conditions

2.3.1. Central location testing

The CLT was performed in DELTA's listening room fulfilling EBU 3276 [8] (this also means that the room conforms to ITU-T BS.1116-1 [1]), meaning low reverberation time (0.25 sec. at most frequencies and 0.5 sec. at the lowest frequencies) and low background noise (below NR10 with ventilation at 75 %).

The projector for displaying the test interface to the assessor is placed in the adjacent control room that is completely decoupled from the listening room to eliminate fan noise. An acoustically transparent projection screen was employed for

[1] Whilst ITU-R BS.1116-1 recommends the usage of 20 expert assessors, extensive testing and experience within SenseLab has shown that stable and repeatable data can be obtained for this size of test with 12 - 15 highly experienced expert assessors.

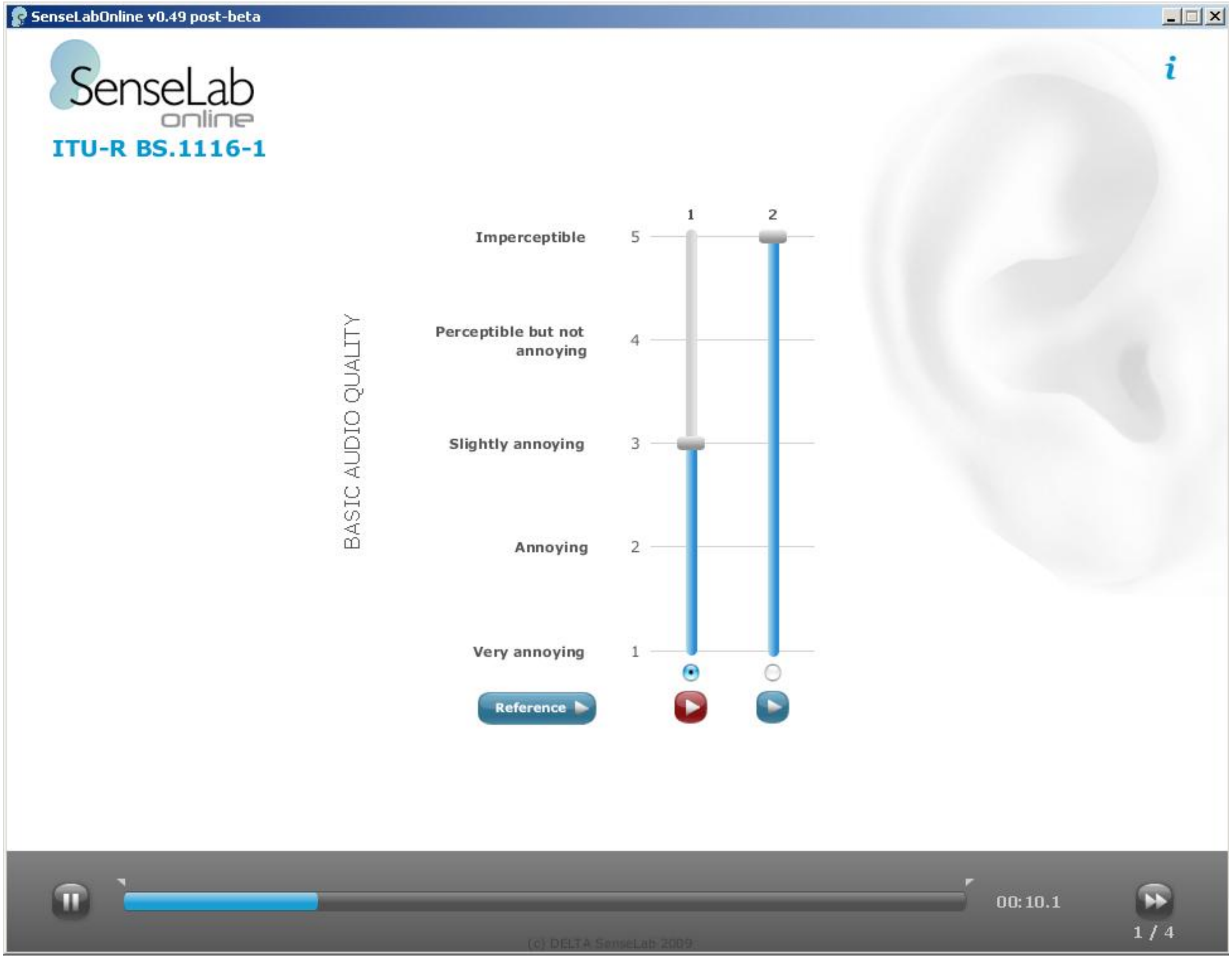

Figure 1: SenseLabOnline [9] web based graphical *user interface as viewed by assessors.*

the presentation of the test software graphical user interface. Computer hardware and other potentially noisy equipment were also located in the control room.

The test stimuli were present via PC via the equipment listed in Table 3 . The balanced output signal from the soundcard was fed to a graphical equalizer and a passive attenuator and from there to the input of at set of active stereo loudspeakers.

Connections between devices in the signal chain were made with professional quality XLR microphone cables.

Device	**Brand**	**Model**
PC	IBM/ Lenovo	ThinkCentre
USB soundcard	Digigram	UAX220v2
Equalizer	Behringer	DEQ2496 UltraCurvePro
Passive attenuator	M-Patch	Blue, 2 ch.
Active loudspeakers	Genelec	8050

Table 3: *List of equipment used for sound reproduction in the central location test.*

A standard stereo loudspeaker configuration as defined in ITU-R BS.1116-1 [1] was employed. The loudspeaker base (distance between loudspeakers center axis) was B = 2.70 meters. Height of loudspeakers (midpoint between tweeter center and woofer center) = 1.1 meter which corresponded to the ear height of a seated listener. The listener position was according to ITU-R BS.1116-1 [1] placed at the distance B from each loudspeaker. The loudspeakers orientation was such that their reference axes passed through the reference position at a height of 1.1 meter. All tests were performed with a single assessor at a time located at the listening position.

Loudspeaker distance from nearest wall was 1.0 meter. The floor was carpeted to minimize the effects of the floor reflections, in accordance with the standard. The loudspeaker setup met the recommendations in ITU-R BS.1116-1 [1].

2.3.2. *Home usage testing*

Device	**Brand**	**Model**
PC/Mac	Not known	Not known
USB soundcard	Sandberg	USB to Sound Link
Headphone	Sennheiser	HD 438

Table 4: *List of equipment used for sound reproduction in the home usage test.*

For the home usage tests, assessors were provided with a moderate quality sound card and headphones as listed in Table 4. They were asked to log into SenseLabOnline using their own computer at home and perform the listening test in quiet conditions whilst focusing upon the assessment task.

The assessors were also instructed to adjust the playback sound pressure level to the most comfortable level during a short training module, where they were presented with a selection of samples from the actual test. They were instructed not to adjust the playback level during the test. The stimulus presentation was performed by the SenseLabOnline system.

3. Analysis

The statistical analysis was performed in 4 steps. In a first step the assessor's performances have been checked in both situations in order to evaluate the reliability of the listeners. The data assumptions for the analysis of variance were subsequently evaluated, followed by an ANOVA for each experiment. Both the overall structure of both datasets was then performed. Finally, a comparative analysis of CLT and HUT was performed to objectively compare the data from the two test situations.

3.1. Assessor performance

In the ITU-R BS.1116-1 [1] there is only one way to have a correct answer: the assessor has degraded the system under test, and he has not degraded the reference.

If the assessor has degraded the reference, the answer is considered as wrong. In the last case, if the assessor has not degraded the reference neither the system under test, the answer is considered as null.

The following graphs represent the percentage of correct answers, each point represent the performances of one assessor for one system.

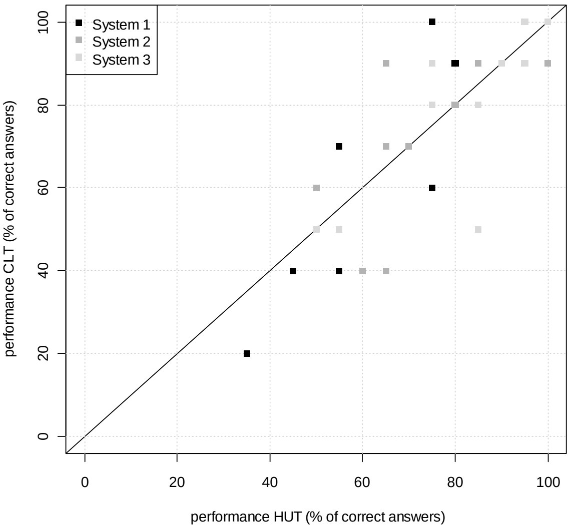

Figure 2: *Assessor detection performance in terms of correct identification of the test item.*

The cloud of points is narrow and centered on the first bisector which means that the performances of the assessors have globally not shifted between the two situations. The most accurate assessors in the CLT are also the most accurate in the HUT. Moreover, the performances under 50% in the CLT have been improved in the HUT. On the contrary, the performances close to 90% in the CLT are a under 90% in the HUT. Thus the performances are globally stable.

3.2. Separated analysis

As said previously, the data collection has been performed with SenseLabOnline [9] and the statistical analysis has been automated with a program in R language [10]. For both data sets the data quality was evaluated prior to performing the analysis of variance and subsequently over viewing the data.

Datasets from both CLT and HUT were reviewed in terms of fulfilling the ANOVA assumptions. In both cases, there is homogeneity of variance, and the residuals (post ANOVA) are also normally distributed and thus an ANOVA model is found suitable to model the data.

3.2.1. Home usage testing

The Figure 3 shows the MOS degradation of the different systems compared to the MOS degradation of the reference. In accordance to the results of the assessor performances, the reference has undergone small and insignificant degradation.

The systems under test were well differentiated from the reference. Nevertheless, the discrimination between systems is not significant. Indeed, the difference between systems 2 and 3 is not significant but the system 1 is significantly different from the others. Moreover, the 95% confidence intervals are narrow thus system 2 and system 3 are considered to be really similar and the assessor discrimination performance is not questioned.

	Sum Sq	**Df**	**F value**	**Pr(>F)**
(Intercept)	403.23	1	2021.63	<2.2e-16
Assessor	90.21	11	41.11	< 2.2e-16
System	20.68	2	51.85	< 2.2e-16
Sample	123.05	9	68.55	< 2.2e-16
Assessor:System	6.38	22	1.45	0.09366
Assessor:Sample	59.53	99	3.01	2.175e-11
System:Sample	5.98	18	1.66	0.04792
Residuals	39.49	198		

Table 5: *ANOVA of the HUT*

The complete model with 2-way interactions has been chosen for the Analysis of Variance (ANOVA), the results are sum up in Table 5. The explained variance (R^2) is 79% of the total variability, hence the analysis is reliable. The ANOVA reveals that the main effects are significant (p-value<0.05) with high F-values.

The interaction Assessor/sample and System/Sample are also significant, but the F-values are low thus they have not as powerful effect as the main effects.

3.2.2. Central location testing

The data collection has been performed with SenseLabOnline, and therefore there are no differences between both tests in the User Interface. The results from the comparison of the degradation of the reference and the different systems under test are sum up in Figure 4.

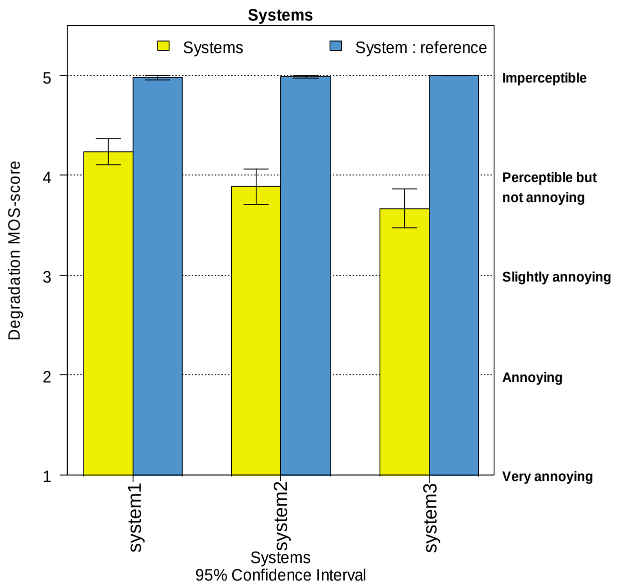

Figure 3: *Combined degradation MOS- score of the systems and the reference (HUT)*

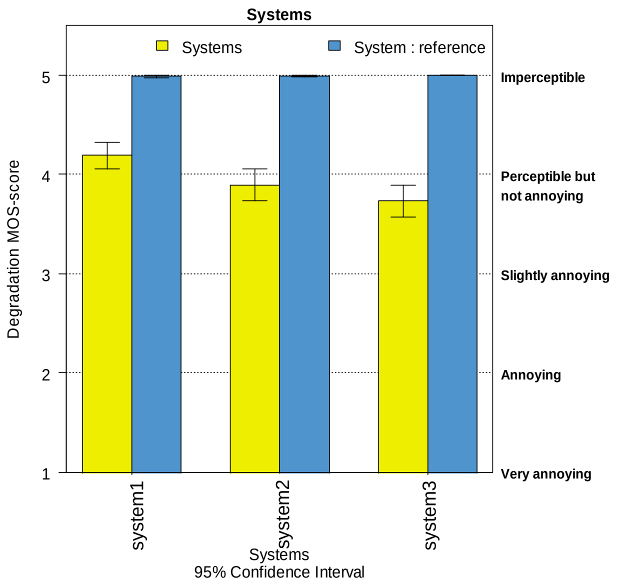

Figure 4 Combined degradation MOS- score of the systems and the reference (CLT)

The complete model with 2-way interactions has been chosen for the Analysis of Variance (ANOVA), the results are sum up in Table 6. The explained variance (R^2) is 74% of the total variability, hence the analysis is reliable. The ANOVA reveals that the main effects are significant (p-value<0.05) with high F-values.

The degradation graphs form the CLT data and the HUT data are really similar, the reference has undergone low degradation (only with the systems 1 and 2) and has been well distinguished from the systems. As in the HUT, the system 1 is significantly different than the systems 2 and 3, but the system 2 is not significantly different than the system 3.

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	40122	1	3249.90	<2e-16
Assessor	3571	11	26.29	< 2e-16
System	1364	2	55.26	< 2e-16
Sample	11145	9	100.3076	< 2e-16
Assessor:System	394	22	1.45	0.095
Assessor:Sample	8087	99	6.6169	< 2e-16
System:Sample	224	18	1.00	0.45
Residuals	2444	198		

Table 6: *ANOVA of the CLT*

The difference in the analysis lies on the ANOVA, where the interaction system/sample is not significant (Table 6). Moreover the F-value of the sample effect is larger in the CLT (100) than in the HUT (68.5). That means that discrimination between samples is bigger in the CLT than in the HUT.

3.3. Combined analysis

In order to have a clearer comparison of both situations (CLT and HUT) both data sets have been merged into one data set.

The linear model chosen for the ANOVA is taking in count the same effect of the previous analysis plus the location factor and its related interaction with the other main factors (Table 7). The results show that the assessor, system and sample effects are significant, but the location effect is non significant. The interpretation is that there is no significant difference in the degradation of the systems under test in the two situations.

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	80444	1	4000.955	< 2e-16
Location	0.1253	1	0.0062	0.93710
Assessor	11005	11	49.7577	< 2e-16
System	3395	2	84.4273	< 2e-16
Sample	22655	9	125.1976	< 2e-16
location:Assessor	1587	11	7.1777	1.883e-11
location:System	38	2	0.9388	0.39175
location:Sample	795	9	4.3927	1.484e-05
Assessor:System	394	22	1.45	0.095
Assessor:Sample	8087	99	6.6169	< 2e-16
System:Sample	224	18	1.00	0.45
Residuals	2444	198		

Table 7: *ANOVA of the merge data from the CLT and HUT*

4. Discussion and conclusions

The analysis of both the CLT and HUT experiments reveals that assessors perform well in the experiments and are able to discriminate the audio codecs in a reliable manner. The ANOVA indicates that there are significant difference to be found between both codecs and program items in both conditions. The explained variance and the contributions (F-value) and significance levels (p-values) from both experiments are comparable. This indicates a degree of similarity between the results of the two experiments.

Upon inspection of the overall data, it can be concluded that there is no statistically significant difference to be found between the means for both experiments. The only difference to be found relates to the ANOVA, where the F-value of the sample effect is wider in the CLT than in the HUT experiment.

Overall this is a very encouraging set of findings especially , when considering some of the detailed aspects of the experiments. Firstly, the CLT was fully calibrated with high quality test equipment, whilst the HUT was performed in a less controlled manner with lower quality equipment. Secondly, the comparison was made between loudspeaker and headphone reproduction. It can be argued whether or not the headphone reproduction is more sensitive than loudspeaker reproduction. Lastly, the interval between the two conditions was quite significant. The first (CLT) test was performed in December 2009 and the second (HUT) in March 2010. Between these two tests assessors only performed unrelated listening tests with different samples and test stimuli.

It can be concluded that this was a very critical small impairment study, with maximum degradations in the order of 4.0 MOS. However, in both experiments expert assessors were able to reliably discriminate these small differences and rate them in a very similar manner leading to identical overall conclusions. This illustrates the potential of home usage testing even for very critical cases and the potential such web based testing can bring in terms of speed, ease of access and opportunity for distributed testing.

5. Acknowledgements

This work was partially performed as part of the HINTLAB and EvaluationLab R&D projects within DELTA SenseLab that were co-funded by The Danish Council for Technology and Innovation - under The Ministry of Science, Technology and Innovation.

6. References

[1] ITU-R, Recommendation BS. 1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems, International Telecommunications Union Radio-communication Assembly, 1997.

[2] S. Bech and N. Zacharov, “ Perceptual Audio Evaluation. Theory, Method and Application”, *John Wiley & Sons, Ltd*, 2006.

[3] I. Boutrollea, J. Delarueb, D. Arranza, M. Rogeauxa and E.P. Köster, Central location test vs. home use test: Contrasting results depending on product type, Food Quality and Preference Volume 18, Issue 3, Pages 490-499, April 2007.

[4] K. Sveinsdóttir, E. Martinsdóttir, F. Thorsdóttir, R. Schelvis, A. Kole, I. Thorsdóttir, Evaluation of farmed code by a trained sensory panel and comsumers in different test settings. Journal of Sensory Science Volume 25, Number 2, pages 280-293, April 2010.

[5] A. Sverkén, A. Åström, K. Wendin, Comparison of Home Use Test (HUT) and Central Location Test (CLT) by the use of Preference Mapping, The 7th Pangborn Sensory Science Symposium, Minneapolis, USA, 2007.

[6] ISO 8586-2. Sensory analysis – General guidance for the selection, training and monitoring of assessors-Part2: Experts, International Organization for Standards 1994.

[7] EBU SQAM CD, Sound Quality Assessment Material - recordings for subjective tests, Compact Disc No 422 204-2, European Broadcast Union, 1988. Also available from http://tech.ebu.ch/publications/sqamcd

[8] EBU 3276-1 Technical document Tech 3286: Supplement 1- Listening conditions for the assessment of sound programme material: monophonic and two-channel stereophonic, European Broadcast Union, May 1998

[9] http://www.senselabonline.com

[10] R CRAN, http://cran.r-project.org/