Sensory evaluation of hearing aid performance based on normal-hearing listeners

Søren Vase Legarth [1], *Christian Stender Simonsen*[2], *Lars Bramsløw*[2], *Guillaume Le Ray*[1], *Nick Zacharov*[1]

[1] DELTA SenseLab, Venlighedsvej 4, DK-2970 Hørsholm, Denmark
[2] Oticon, Kongebakken 9, DK-2765 Smørum, Denmark
svl@delta.dk, css@oticon.dk

Abstract

This paper describes a method used for evaluating sound quality of 4 hearing aid products using expert assessors for a sensory profiling test and a preference test.

Hearing aids were evaluated by trained normal-hearing listeners in 7 different sound environments that were selected to represent some of the common hearing aid user situations. The scenarios were reproduced in a standard listening room. Assessors performed a consensus language development to establish seven key sound quality attributes, which they used to assess the performance of each hearing aid in each sound scenario. Additionally, preference data was also collected from assessors. The univariate analysis of the preference data and the multivariate analysis of the sensory profiling data are discussed in detail.

1. Introduction

Traditionally hearing aids are designed to improve speech intelligibility for hearing impaired persons. A range of well described and standardised speech intelligibility test methods like HINT (Nilsson et al, 1994), SPIN (Bilger et al, 1984), Hagerman (Hagerman, 1982 and Wagener, 2003) and numerous others exist and are widely used in the hearing aid industry [1], [2], [3] and [4].

Speech intelligibility is the most important factor when evaluating hearing aids but sound quality has also been found to be an important parameter. Bentler et al (1993) found sound quality to account for 20 % variance in hearing aid satisfaction [5]. As the deployment of hearing aids shifts towards a younger population with mild hearing losses, it appears that the matter of sound quality in hearing aids becomes more prevalent.

Testing subjective measures like sound quality demand for well thought out test designs in order to make sure results are reproducible and reflect everyday listening situations. Standardised questionnaires like the Speech, Spatial and Qualities of Hearing Scale (SSQ) (Gatehouse & Noble, 2004) are extensively used in field experiments [6]. But the inherent problem of testing several algorithms or different hearing aids can be tiresome and quite demanding for the test person in real life situations and also consumes a fair amount of time. Thus, more advanced sound quality evaluation methods are required.

Gabrielsson et al (1988) developed the Judgement of Sound Quality (JSQ) to evaluate sound quality of systems [7]. It employs eight dimensions of sound quality: spaciousness, loudness, softness, clarity, fullness, nearness, brightness and an overall fidelity. Narendran & Humes (2003) concluded the method to be a potentially useful measure of hearing aid performance, but also noted that some effort should be put into improving its' reliability [8].

An alternative approach to study the perceived sound quality in hearing aids can be based upon sensory evaluation techniques, as developed in the food industry and thoroughly described in [9]. A state of the art review of the field and its deployement in sound and acoustic application can be found in [10].

This paper describes the method used for evaluating the perceived sound quality characteristics of 4 high-end hearing aids from different manufacturers. The method was demonstrated using trained normal hearing assessors and with the hearing aid prescribed for a mild standardized hearing loss N2 defined in IEC 60118-15 draft version [11].

2. Methods

The sound quality of the hearing aids was approached from two angles: Sensory profiling and overall preference. The sensory profiling was performed with a normal hearing expert assessor panel to define and assess the primary attributes that characterized the perceived sound quality of the hearing aids. The preference test was performed with normal hearing listeners.

Since the overall project purpose was to evaluate the validity of the methods in relation to sound quality assessments, the use of normal hearing assessors seems justified.

But to collect valid results for future hearing aid product development, it would be an advantage to employ real hearing aid end-users (at least) to assess the overall preference.

2.1. Consensus attributes development

First part of the sensory profiling process was to familiarize the assessors with the range of hearing aid products. Ten assessors from DELTA SenseLab's trained assessor panel were employed for the task.

For the purpose of identifying 6-8 relevant attributes for the sensory profiling, the assessors were given the task to listen to HATS recordings of hearing aids in selected sound scenarios and write down a list of relevant attributes (individual word elicitation process). The individual lists from all 10 assessors were reviewed by the panel leader who condensed it into a list of the primary eight attributes including suggested anchor labels and definitions (one attribute was removed later in the process). This list was the initial working document to be used in the plenum session with the panel.

10.21437/PQS.2010-25

The main work of the plenum session was to reach consensus on the attributes and define the anchor labels and unambiguous definitions.

During the consensus attributes development session, the assessors listened via headphones to the calibrated HATS recordings of the products. In this way the attributes, anchors and definitions were pre-tested to prove their significance. During the process it was emphasized by the panel leader that the sensory profiling test would be a Semantic Differentials test type which sets high requirements to precise definitions and highly discriminative attributes.

The list of the final seven attributes is given in Table 1.

Attribute list	Lower anchor	Higher anchor
'Tube'-sound	Faint	Clear
Sharpness	Little	Much
Loudness	Soft	Loud
Inherent noise	Faint	Clear
Overload	Little	Moderate
Room-perception	Damped	Hard
Source reproduction re. Reference	Recognizable	Unrecognizable

Table 1. *List of attributes defined by the panel in the plenum session to be used for hearing aid sensory profiling.*

2.2. Attributes assessments

A Semantic differential type of method was applied for the attribute assessments. The Semantic differential method was originally developed by Osgood (1957) and designed to measure the connotative meaning of concepts [12]. In this project the method was modified from its original definition to assess the attributes for each product/scenario combination instead of using adjective pairs. Each attribute was represented by a scale and corresponding anchor labels. This way of assessing all attributes in each trial was chosen as the best compromise for assessing real products that includes physical handling (taking hearing aid on/off). The down-side of the method is that it puts high demands on the assessors acoustical memory due to the lack of possibility to directly compare between products.

In order to increase the product discrimination, a reference was included. The reference was available for the assessor during the assessments and was predefined to represent a given scale rating for each attribute.

A balanced block design was applied for the product presentation order on assessor level. All products were evaluated for one sound scenario at a time.

2.3. Preference assessments

The preference test was based on the ITU-R BS.1534-1 (MUSHRA) recommendation [13] (a "double-blind multiple-stimulus with hidden reference and anchor" method). The reference samples and the anchor samples should span the range of program items and artefacts to be expected in the listening test.

Stimuli were rated according to the continuous quality scale which is divided into five equal intervals labeled: 'Bad', 'Poor', 'Fair', 'Good' and 'Excellent'.

The overall attribute evaluated in the test was *Basic audio quality*.

For all assessors a randomized presentation order was used within each repetition block, yielding the double blind paradigm.

3. Experimental setup

3.1. Listening panel

10 assessors from DELTA's selected assessor panel were employed for the consensus attribute development and sensory profiling. Eight assessors were males and 2 assessors were females. The age was 24 – 48 years with an average of 33 years.

The assessors were all native Danish speakers with normal hearing according to ISO 8253-1 [14] and had qualified for the selected assessor panel through various discrimination tests on small audio impairments as described by Legarth & Zacharov (2009) in [15].

Prior to this project the selected assessors had received training and familiarization in similar tasks.

15 assessors from DELTA's selected assessor panel were employed for the MUSHRA preference test. 13 assessors were males and 2 assessors were females. The age was 22 – 48 years with an average of 31 years.

3.2. Listening room and loudspeaker setup

DELTA's listening room fulfils EBU 3276 [16] (which also means that the room conforms to ITU-T BS.1116-1) [17], meaning low reverberation time (0.25 sec. at most frequencies and 0.5 sec. at the lowest frequencies) and low background noise (lower than NR10 with ventilation at 75%).

For the consensus attribute development work, the listening room was furnished with one large table and headphones were provided for all assessors.

For the attributes assessments, the listening room was furnished in a calibrated 5.0 multichannel loudspeaker setup according to ITU-R BS.1116-1 [17].

The reproduction level of the selected sound scenarios was adjusted to a realistic level. This level was also used for the recordings of all the hearing aids in the sound scenarios to be used in the MUSHRA preference test.

Sound Scenario	L_{Aeq} (dB)
Office	59.8
Traffic	71.7
Water	60.6
Forest	55.8
Females talking	61.8
Speech	56.1
Music	65.7

Table 2. *A-weighted sound pressure levels of the sound scenarios measured at listeners position.*

3.3. Stimuli and presentation

Four different hearing aids were included for the test. Three of the hearing aids were high-end RITE-type (Receiver In the Ear) products from 3 different manufacturers and one was a BTE-type (Behind the Ear).

The hearing aids were all programmed with the N2 prescription which is targeting mild hearing losses and fitted individually to each assessor.

Seven different sound scenarios were chosen for testing the hearing aids. The scenarios were field recordings of everyday situations recorded with a Soundfield microphone in B-format and decoded with a digital surround processor into 5.1 multichannel format.
The 7 sound scenarios were described as:

1. Office (open office environment with people talking and printers making noise)
2. Traffic (road traffic noise from a street corner with scooters, cars, motorcycles and a truck)
3. Water (a small spring in the forest)
4. Forest (bird song and wind picking up in the tree tops)
5. Females talking (talk in a meeting break with 10+ women in a moderately reverberant room)
6. Speech (Danish female speaker).
7. Music (Jazz song performed by Diana Krall: 'Let's fall in love', vocal and piano solo).

The scenarios were selected to be representative for different hearing aid user environments but also selected as good stimuli for evaluating different aspects of sound quality.

The data collection was performed with DELTA developed Labview software: A screenshot of the software user interface is shown in Figure 1.

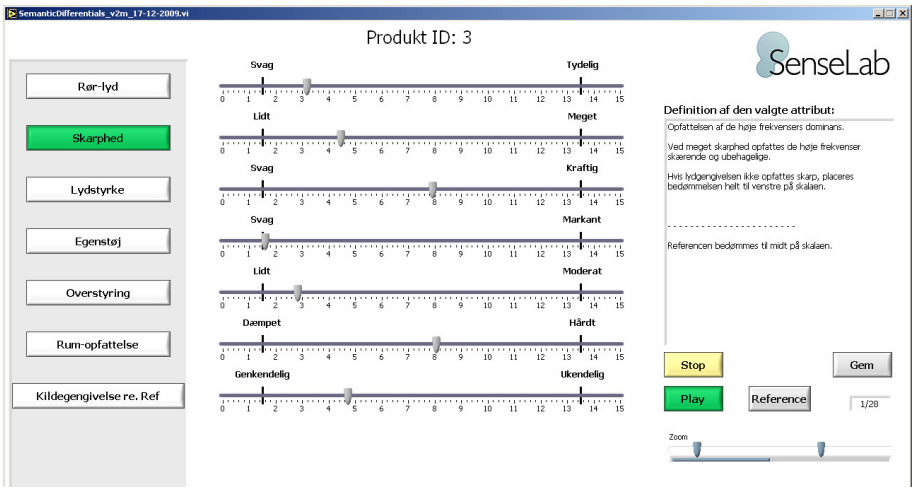

Figure 1: *Screenshot of Graphical User Interface for collecting attribute assessments of the hearing aids.*

In the top of the screen the assessor can see what product /hearing aid should be evaluated. When the assessor clicks on an attribute (turns green), the corresponding definition is displayed in the right side of the screen. The definition also includes the scale value for the reference stimulus.

The test software controls the playback of the multichannel sound sample and the assessor is able to zoom into a specific section of the scenario.

On completion of a product assessment, the assessor presses the 'Gem' (Save) button and the software checks if all attributes have been assessed before continuing to next product. In case of missing data, the assessor is receiving a message on the screen.

4. Statistical analysis and results

4.1. Assessor performance for attribute ratings

This part of the data analysis focuses on the reproducibility and quality of the attribute assessments. The panel performance is a measure for how well data can be trusted. If the panel is behaving inconsistently, the data will be noisy and less valid.

The first point of interest is to check for the panel repeatability. The test included two replicates and the data has been submitted to an analysis of variance (ANOVA). The results given in Table 3 shows, that the products are significantly different except for 'Tube'-sound. There is no significant difference between the assessments of the products for the two repetitions. It is also seen that there is a significant assessor effect but this is to be expected in sensory evaluations of products.

P-value	System	Assessor	Sample	Replicate
Inherent noise	**1.42E-06**	**7.17E-55**	**5.63E-12**	0.342
Loudness	**0.000459**	**9.76E-92**	0.41	0.815
Overload	**0.00788**	**1.56E-74**	**0.000367**	0.899
Room-perception	**0.00581**	**5.73E-51**	0.0627	0.64
Sharpness	**0.000266**	**1.68E-24**	0.0534	0.23
Source reproduction	**1.01E-05**	**6.27E-60**	**4.57E-10**	0.708
'Tube'-sound	0.717	**5.71E-68**	**4.90E-11**	0.128

Table 3. *ANOVA (ANalysis Of VAriance) made on both replicates for the attribute assessments. P-values less than 0.05 are bold to indicate significant levels.*

F-value	System	Assessor	Sample	Replicate
Inherent noise	**75.6**	**43.4**	**164**	**4.14**
Loudness	**153**	**88.8**	**2.22**	0.181
Overload	**28.4**	**65.4**	**26.5**	0.0303
Room-perception	**13.1**	**39.6**	**6.83**	0.849
Sharpness	**37.3**	**17.4**	**7.06**	**5.54**
Source reproduction	**36**	**48.7**	**27.8**	0.342
'Tube'-sound	**3.19**	**57.5**	**39.7**	**3.95**

Table 4. *ANOVA (ANalysis Of VAriance) made on both replicates for the attribute assessments. F-values greater than 2 are in bold.*

From the ANOVA it is found that the products (System) are significantly different for all attributes except Tube'-sound which means that they are well discriminated. The sample effect is significant for the attributes: Inherent noise, Source reproduction re. Reference, Overload and 'Tube'-sound. For the remaining attributes the samples are not assessed significantly different across sound scenarios.
The interaction between product and sample is also of interest as it tells whether the products are perceived differently across sound scenarios. This is the case for all attributes except Room perception. The interaction between product (system) and sample is expected since the sound scenarios were chosen for their diversity.

4.2. Product performance

In this analysis, the average value across repetitions is used, since the ANOVA showed that there was no significant difference between replicates.

One way of representing the results from the hearing aid assessments is to use profile plots. The plots are showing the panel average rating for each attribute on the 4 different products for each sound scenario. In the profile plots in Figure 2 to Figure 4 the 4 hearing aids are shown for the sound scenarios: Water, Speech and Music. These plots are used for identifying the general sound scenario profile and how the products differentiate from each other.

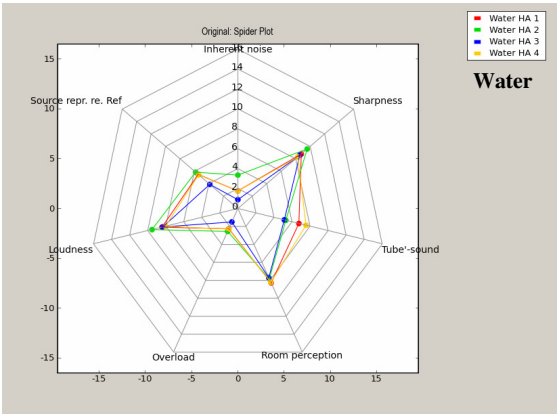

Figure 2. *Profile plot showing the average rating of each attribute and hearing aid for the sound scenario: Water.*

The profile plots show an overall difference between the scenarios Water and Speech + Music. The Water scenario does not expose the Inherent noise in the hearing aids as well compared to the Speech and Music scenarios, but 'Tube'-sound is more dominant in the Water scenario. It is also seen that the attribute Source reproduction is discriminating the products differently in the scenarios.

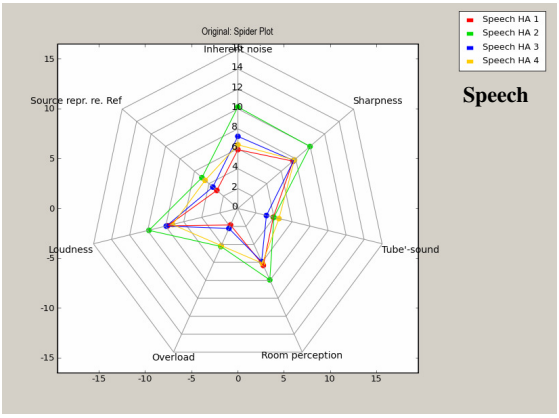

Figure 3. *Profile plot showing the average rating of each attribute and hearing aid for the sound scenario: Speech.*

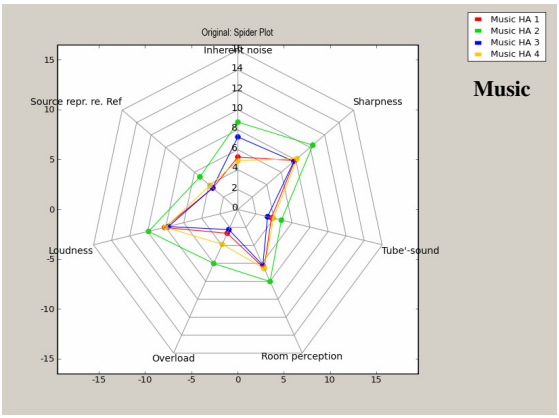

Figure 4. *Profile plot showing the average rating of each attribute and hearing aid for the sound scenario: Music.*

4.3. Multivariate analysis

A Hierarchical Multiple Factor Analysis (HMFA) was performed on the data set to simplify the interpretation of the main product differences [18]. Data was normalised for this analysis to improve the resolution in the data set. The global analysis yielded a complex result due to the highly adaptive processing in the hearing aids and the large differences between the scenarios.

To obtain a more simple interpretation of the data, a hierarchical clustering was performed to identify the families of sound scenarios that produced the same relative perception of the hearing aids. The clustering resulted in a Dendrogram shown in Figure 5.

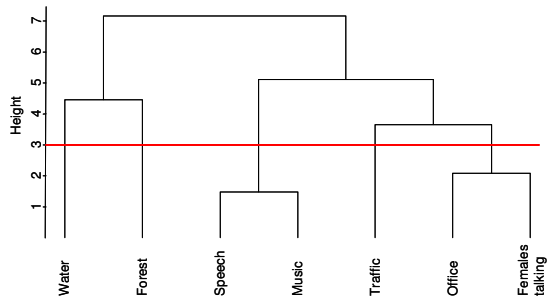

Figure 5. *Cluster dendrogram showing the related sound scenarios.*

From the cluster dendrogram in Figure 5 it is seen, that the sound scenarios can be categorised into 5 groups:

1. Water (dynamic environmental sound)
2. Forest (stationary environmental sound)
3. Speech, Music (clean studio recordings with large voice to noise ratio)
4. Traffic (dynamic ambient sound with large amplitude variation)
5. Office, Females talking (dynamic ambient sound with intelligible speech)

Each of the groups were then analysed using HMFA but for the purpose of illustrating the methodology only the results from group 1 and group 3 will be addressed here.

The HMFA is based on principal component analysis for each sound scenario. By mapping each data set onto a common component structure, information on the main characteristics describing the differences of the hearing aids are found.

In Figure 6 and Figure 7 the Bi-plots from the HMFA are showing the 95% confidence ellipses and the attribute loadings on the dimensions. The explained variance for each dimension is given. The high explained variability in the data indicates that the products are clearly different for one or more attributes.

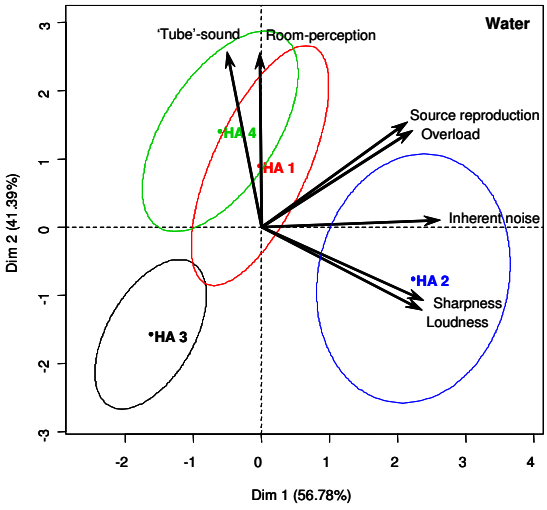

Figure 6. *Bi-plot from the HMFA of the Water scenario. The ellipses indicate the 95% confidence level. The first dimension is explained by Inherent noise. The second dimension is explained by 'Tube'-sound where HA 3 is the best at suppressing this negative effect. HA 4 is rated to have most 'Tube'-sound for this scenario.*

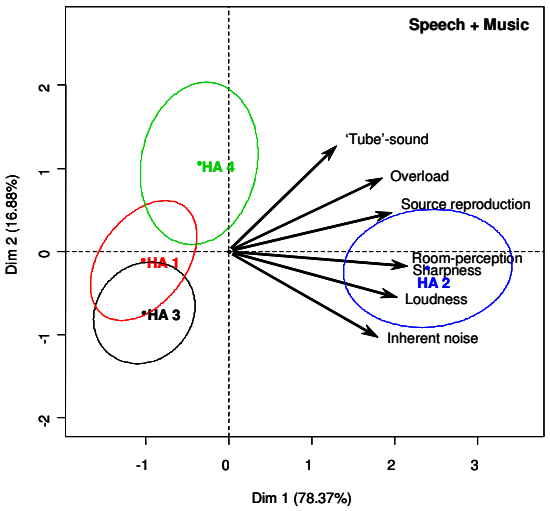

Figure 7. *Bi-plot from the HMFA of the Speech and Music scenarios. The ellipses indicate the 95% confidence level. In this group of scenarios, HA 2 is characterized as a product with many negative sound quality features like: Sharpness, Overload and Inherent noise.*

The confidence ellipses in Figure 6 and Figure 7 show that HA 2 is clearly different from the other 3 hearing aids. To identify in which way HA 2 is different one have to look at the attribute vectors. It is seen that HA 2 mainly is perceived as having more Loudness, Sharpness and Inherent noise than the remaining hearing aids. This is complementing the findings from the profile plots. HA 3 is characterized by little Overload and 'Tube'-sound and is the better product to reproduce the Water scenario as a recognizable sound.

HA 1 and HA 4 are middle range products with more or less the same characteristics, though HA 4 seems to have a bit more 'Tube'-sound.

4.4. Overall preference

The ITU-R BS.1534-1 (MUSHRA) [13] test is designed to detect and quantify differences between the systems under test (hearing aids). Figure 8 show the mean opinion scores of the assessor ratings for all samples (sound scenarios).

The results show that there is a significant difference between the hearing aids. HA3 is perceived to have a significant better sound quality than the three other hearing aids. HA2 is rated to have the significantly worst sound quality.

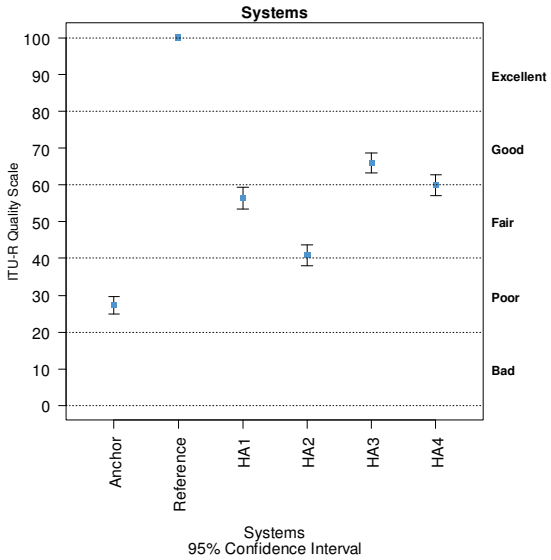

Figure 8. *Mean Opinion Scores of the overall preference test indicated with 95% confidence intervals.*

By combining the results from the sensory profiling with the overall preference it is possible to identify the characteristics (attributes) that are directly related to the perceived sound quality of the hearing aids.

5. Conclusions

A method has been described for evaluating the perceived sound quality of hearing aids.

A panel of 10 trained native Danish assessors has worked towards the development of 7 attributes for assessing the perceived sound quality of hearing aids. The final attributes are presented in this paper.

The attributes have been developed in a plenum session where the assessors as preparation had performed individual word elicitation on HATS recordings of the products and sound scenarios.

The performance of the assessors in attribute rating was good and data was reproducible.

Four different hearing aids were evaluated for the 7 attributes in 7 different sound scenarios. The scenarios were reproduced in a calibrated multichannel loudspeaker setup and the assessors evaluated the real hearing aid products in a Semantic Differentials test design.

Initially a ***global*** analysis was performed to look at the performance of the main factors (assessor, product, repetitions, and sound scenario. The results shows, that the products are significantly different and that there is no significant difference between the assessments of the products across repetitions.

The ***profile plots*** allow for an overview of the performance of product and scenario interaction, providing an overall impression of the variation for each case. In general we can observe a range of characteristics for each product. However, from this global level, a limited amount of beneficial information can be extracted.

The next level of analysis was to look at the data from a ***multivariate*** perspective and to dig into the complex interactions using the hierarchical multiple factor analysis technique. The results showed that beneficial information can be extracted from the analysis to identify the main characteristics to describe the differences between the hearing aid products.

6. Acknowledgements

The INTERSPEECH 2010 organizing committee would like to thank the organizing committee of INTERSPEECH 2009 for their help and for kindly providing the template files.

7. References

[1] Nilsson, M., Soli, S. D., Sullivan, J. A. (1994), *Development of the Hearing in Noise test for the Measurement of Speech Reception Thresholds in Quiet and in Noise*, Journal of the Acoustical Society of America, Vol. 95 (2), pp. 1085-1099.

[2] Bilger, R. C., Nuetzel, j. M., Rabinowitz, W. M., Rzeczkowski, C. (1984) *Standardization of a Test of Speech Perception in Noise*, Journal of Speech and Hearing Research, Vol. 27 (1), pp. 32-48.

[3] Hagerman, B. (1982), *Sentences for Speech Intelligibility in Noise*, Scandinavian Audiology, Vol. 11, pp. *79-87.*

[4] Wagener, K., Josvassen, J. L., Ardenkjaer, R. *(2003), Design, Optimization and Evaluation of a Danish Sentence Test in Noise, International Journal of Audiology*, Vol. 42, pp. 10-17.

[5] Bentler, R. A., Niebuhr, D. P., Getta, J. P., Anderson, C. V. (1993), *Longitudinal Study of Hearing Aid Effectiveness. I: Objective Measures*, Journal of Speech and Hearing Research, Vol. 36, pp. 808-819.

[6] Gatehouse, S., Noble, W. (2004), *The Speech, Spatial and Qualities of Hearing Scale (SSQ)*, International Journal of Audiology, Vol. 43, pp. 85-99.

[7] Gabrielsson, A., Schenkman, B. N., Hagerman, B. (1988) *The Effects of Different Frequency Responses on Sound Quality Judgements and Speech Intelligibility*, Journal of Speech and Hearing Research, Vol. 31.

[8] Narendran, M. M., Humes, L. E. (2003) *Reliability and Validity of Judgments of Sound Quality in Elderly Hearing Aid Wearers*, Ear and Hearing, Vol. 24 (1), pp. 4-11.

[9] Lawless, H. T., Heymann, H., *Sensory Evaluation of Food – Principles and practices*, Springer, 1998

[10] Lorho, G., *Perceived quality evaluation - an application to sound reproduction over headphones*, PhD thesis, Aalto University School of Science and Technology, Espoo Finland, June 2010.

[11] IEC 60118-15 1WD, *Electroacoustics – Hearing aids – Part 15: Signal processing in hearing aids*, 2008

[12] Osgood, C. E., Suci, G. J., Tannenbaum, P. H. *Measurement of Meaning*, University of Illinois Press, 1957.

[13] ITU-R Recommendation BS .1534-1, *Method for the subjective assessment of intermediate quality level of coding systems*, International Telecommunications Union Radiocommunication Assembly, 2003.

[14] ISO 8253-1, *Acoustics – Audiometric methods – Part 1: Basic pure tone air and bone conduction threshold audiometry*, International Organisation for Standards, 1989.

[15] Legarth, S. V., Zacharov, N. (2009), *Assessor selection process for multisensory applications*, In proceedings of the 126th Convention,Munich, Germany.

[16] EBU 3276, *Listening conditions for the assessment of sound programme material: monophonic and two-channel stereophonic*, Geneva, Switzerland, 1998

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

[18] Pagès, J. (2005), *Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley*. Food Quality and Preference (16), 642-649.