CONCERT HALL ACOUSTICS ASSESSMENT WITH SENSORY EVALUATION – TOOLS AND PRACTICES

T Lokki Aalto University School of Science, Dept. of Media Technology, Finland
J Pätynen Aalto University School of Science, Dept. of Media Technology, Finland
N Zacharov Senselab, DELTA, Denmark

1 INTRODUCTION

Sensory evaluation methods have been predominantly developed in food and wine industry to explore the perceptual characteristics of products, which are hard to evaluate through consumer based preference methods due to huge variation in individual tastes.

The acoustics of concert halls is also heavily influenced by a matter of taste. Therefore, sensory evaluation methods are very useful for studying auditorium acoustics due to their ability to extract information often hidden behind preference judgements. With such methods the sensory profiles of concert halls or profiles of seats inside one concert hall can be formed. Lorho[1] has presented a classification of measurement methods which position the sensory evaluation methods with regard to the well know traditional methods, see Fig. 1. The classification can be adapted to acoustics evaluation as follows. In a concert hall the stimulus is symphonic music or an impulse response. The measurement devices are microphones and microphone arrays or the human listeners. In the physical domain, the room acoustical parameters are derived from impulse responses to give highly objective results. On the other hand, in the affective domain preference judgements might give an overall average picture, but the variance in the data is typically large due to the differences in personal taste and previous experiences of the assessors. Sensory evaluation methods provide a link between these domains enabling profiling of the halls with perceptual characteristics. Such profiles are useful to interpret physical measurement data and can help to explain the preference ratings.

1.1 The sensory evaluation methods suitable for concert hall studies

A range of different methods exist for sensory evaluation in the food and wine industry[2]. Some of those methods have been adapted for audio and acoustics studies, see recent examples in[1,3,4,5]. This paper concentrates particularly to the *individual vocabulary profiling (IVP)* based methods[1,6,7], in which the assessors first develop their own attributes and then use these attributes to provide ratings. The basic premise for the IVP approach assumes that there exist common salient characteristics that will be perceived by assessors in a *similar* manner. Using multivariate statistical analysis techniques, it is possible to extract the common underlying multidimensional perceptual space, which can then be interpreted through the usage of the individual attributes.

Sensory evaluation can also be performed using *consensus vocabulary profiling (CVP)* where a group of assessors first elicit the adjectives to describe the stimuli and then with group discussions develop a common vocabulary of consensus attributes. CVP approach represents one of the most common tools in sensory science, but is very challenging to tailor to concert hall acoustics studies due to physical constraints. Such process would require the initial development of a consensus language based on visits to concert halls with live orchestras, followed by a second round of visits for attribute ratings. The process would thus be very laborious and would also require expert listeners. To be able to do such studies in the future we are first trying to understand the primary consensus attributes with IVP studies. Later, it might be reasonable to have an expert panel to develop the consensus language.

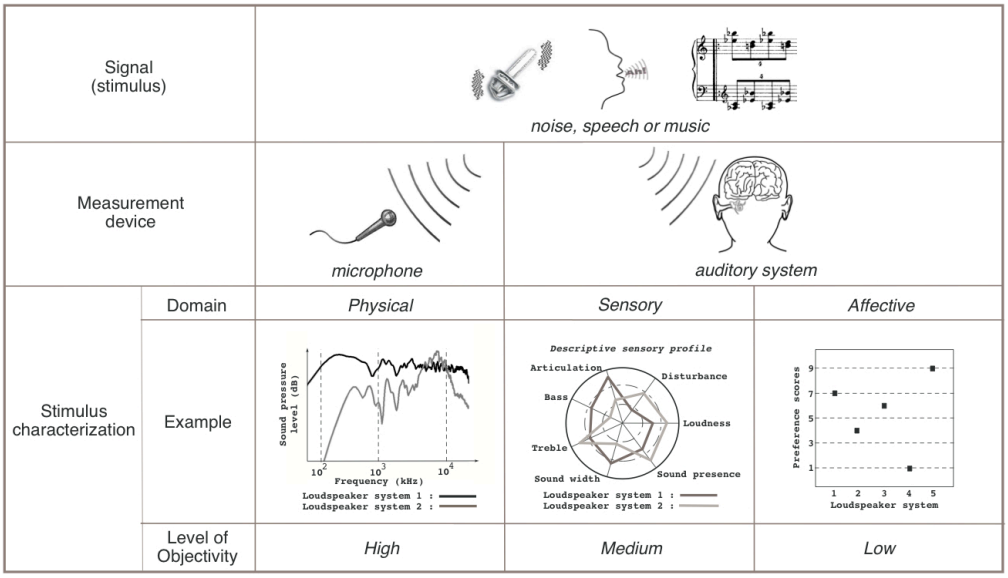

Figure 1 Classification of measurement methods in the physical, sensory and affective domains, as presented by Lorho[1].

The IVP approach allows the assessor to employ their own attributes and thus overcomes the need for assessors to interpret the complex means of consensus attributes used in CVP. For example, the attribute *clarity* can have several definitions and it cannot be ensured that all assessors would understand and agree upon the meaning and usage of a such an attribute scale. The results of individual vocabulary development studies can give a lot of detailed information of the perception of acoustics, in addition to the method being relatively rapid to implement. The discriminating attributes elicited by the assessors provide valuable information as such, but the ordering of samples with these attributes in a common factorial space enables us to create sensory profiles of the studied concert halls[8].

This paper presents tools and practices for concert hall acoustics studies with IVP methods. In particular, the requirements for the sound signals and the testing of the reliability of the assessors are discussed. In addition, an example data set is analyzed by explaining the steps in the analysis. The data set and the function calls for advanced statistical tools are available at `http://auralization.tkk.fi/sensory`.

2 ASSESSOR SELECTION AND TRAINING

A sensory evaluation process is more laborious than, e.g., an affective preference test, primarily due to the large number of attributes to be rated by each assessor. Typical evaluation process is presented in Fig. 2. First, the assessors have to be selected with careful screening in the interest of good data quality[1]. The *selected assessors*[9] familiarize themselves with the samples and elicit attributes on the perceived differences between samples. After some training and definition of attributes, the assessors are ready for evaluation which should consist of at least one rehearsal session.

In general when performing sensory evaluations, it is beneficial to select assessors with care to ensure the quality of collected data. The suitability of assessors is typically reviewed in terms of their *discrimination* ability and *reliability* as discussed in[10,11]. The assessors do not need to be experts in concert hall acoustics nor classical music. It is more important that the assessors can hear differ-

[1] Due to individual nature of the IVP approach, it maybe also be applied with naïve[9] assessors or consumers.

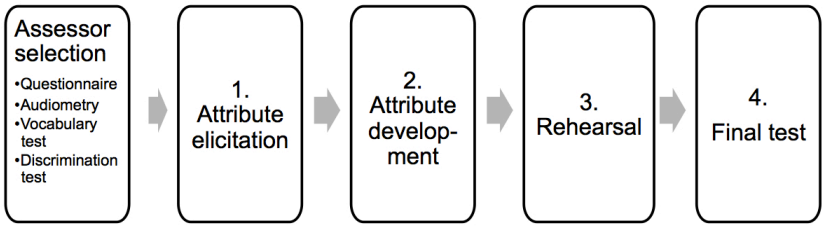

Figure 2 The process of sensory evaluation with individually elicited attributes.

ences between samples and can verbalize well what they hear. In our experience, however, people who often go to concerts and actively listen to recordings are good candidates. Musical background is not needed, but naturally musically trained people are more interested in such evaluation and they probably have better motivation. In addition, they are trained to listen carefully and pick up differences in sound signals.

Typically, the assessor selection process includes at least standard audiometry and some kind of discrimination test. Naturally, the assessors should not have significant hearing loss at any frequencies. The discrimination test can be performed with any methods, but a good convention is to use ABX paradigm[3]. One good practice is to use the same samples in discrimination test as is used in the actual sensory evaluation. This helps the assessors to familiarize themselves with the samples. The discrimination test can even be considered as the first phase in the evaluation, because the assessors can already make notes about the perceived differences. That helps later to define the perceptual differences between all samples. More detailed discussion about assessor screening are presented by Mattila et al.[12], Wickelmaier et al.[13], Legarth et al.[14] and Kuusinen et al.[15].

3 STIMULI, USER INTERFACE, AND LISTENING SPACE

The key point in sensory evaluation is the comparison of samples with descriptive attributes. Ideally, the best data is obtained when direct comparison of samples can be performed. For concert hall acoustics this is challenging to achieve as the assessors can not jump from seat to seat or from hall to hall at the blink of an eye. Therefore, the concert halls have to be recorded for comparison in the laboratory condition. Kürer at al.[16] and Schroeder et al.[17] were among the first researchers who made the instant comparison of concert halls possible by applying binaural technology. In addition, Schroeder et al.[17] enabled the comparison of halls with spatial sound reproduction in laboratory conditions by exciting halls with anechoic recordings, played back by two loudspeakers on the stage.

We have extended this solutions to simulate the whole symphony orchestra by using a loudspeaker orchestra[18] that can be calibrated and which occupies the stage as a real orchestra. For capturing the spatial impulse responses from every single loudspeaker on the stage we use a 3D intensity probe, having three microphone pairs in orthogonal axis. Such technique enables the state-of-the-art spatial sound rendering in the laboratory with Spatial Impulse Response Rendering[19,20]. Furthermore, we use the anechoic symphony orchestra recordings that are publicly available[21]. Even though this is a quite complex way to record concert halls, it guarantees the calibrated orchestra which plays exactly the same music in every hall with the same level. This is very important because only then the subtle differences between halls can be found.

Large differences between samples most probably give results with most obvious perceptual differences, i.e., small differences might be left hidden. In concert halls such overriding characteristics are, e.g., loudness and reverberance. For example, having three seats from three halls might not give the best results between halls as the difference in loudness is large between front and back row seats[8].

In the laboratory the listening space should be quiet, but anechoic chamber is not required. A well

treated listening room with good quality loudspeakers is considered sufficient. The loudspeaker setup should reproduce spatial sound at least from the sides and above. Currently, we are using 14 channel setup; 8 loudspeakers in the horizontal plane, 4 loudspeakers in 45 degree elevation and two frontal speakers below the ear level. This setup covers the directions where the sound reaches the listener in most of the concert halls. The alternative use of headphones for reproduction is tricky as the headphones should be individually compensated for all listeners and head rotations should be made possible with head-tracking, some solutions are described in Spikofski et al.[22] and Algazi et al.[23].

The test user interface was developed using the graphical programming environment MAX 5, to allow for the instantaneous switching between stimuli that are subsequently rated by each assessor. For a given trial, one music sample is presented for comparison in all of the concert halls and evaluated for a randomly selected attribute on a 120 point continuous unstructured line scale. Stimuli are presented double blind.

4 CASE STUDY WITH SIMULATED CONCERT HALLS

Analysis of IVP data is a multi-phase process. Here, an example is given with the real data obtained from our recent research[24] in which the properties of early reflections in a simulated concert hall were studied. First, the sound samples and listening test implementation are briefly revised. Then, the process of data analysis is explained. The analysis and visualizations are done with R, the open source statistical software (`http://www.r-project.org/`). The data and function calls are available at `http://auralization.tkk.fi/sensory` to encourage people to perform IVP studies and similar data analysis.

4.1 Sound samples and motivation for the study

The example data set is obtained by comparing six artificial concert halls. They were created by simulating a symphony orchestra with 24 source positions and computing from each of them the direct sound, 11 early reflections and the late reverberation. All six halls had the same direct sounds and late reverberation, faded in between 60 and 120 ms after the direct sounds. The variation between halls were in the early reflections which were simulated with the image source method from 11 surfaces. The simulated concert halls had three types early reflections which reach the listener even from side (M1, M3, M5) or from close to the median plane (M2, M4, M6). The types of reflections were as follows:

- Concert halls M1 and M2 had 11 reflections from the hard flat surfaces. Such a specular reflection does not violate the temporal envelope of sound.
- Concert halls M3 and M4 had 11 reflections from six different type of diffusors. The responses of diffusors were measured in a semi anechoic space with six different diffusing structures on top of a hard surface. As the measured structures introduced high frequency attenuation, the attenuated energy was compensated by adding 6 ms of spectrally shaped noise 3 ms after a reflection. Together, the measured reflection and the compensation noise had an average flat frequency response, but the temporal envelopes of unresolved harmonics at high frequencies are more or less scrambled.
- Concert halls M5 and M6 had 11 artificial reflections, which were obtained by spreading the energy of a specular reflection to 10 ms time span. This was performed by producing a 10 ms long noise burst with an average flat frequency response. Such a reflection distorts the temporal envelope of sound at all frequencies.

It is important to notice that the total sound energy remains unchanged in all of the six artificial halls (M1-M6), resulting in the same standardized ISO 3382-1 monaural room acoustical parameter values, as seen in Fig. 3. Lateral energy fraction was the same in (M1, M3, M5) and in (M2, M4, M6), respectively. Fig. 3 also illustrates the different types of early reflections with spectrograms.

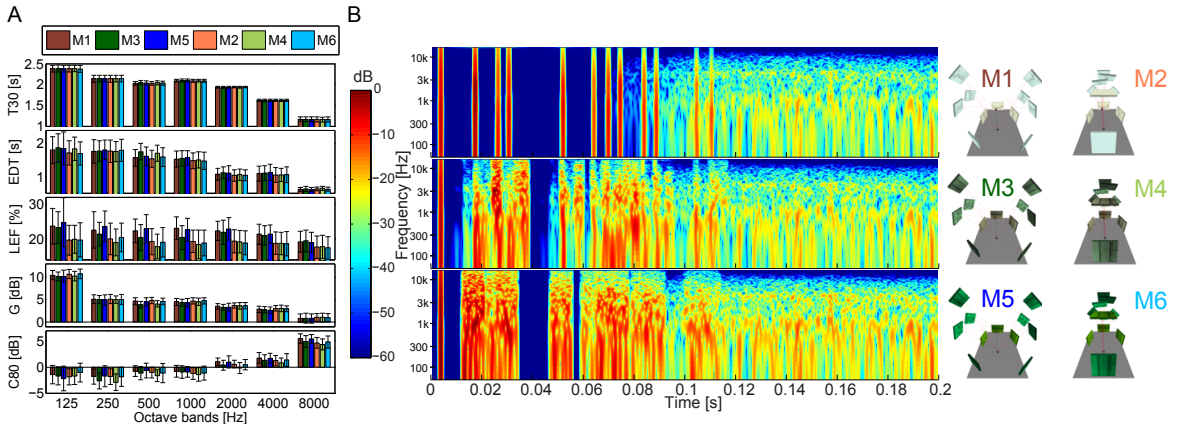

Figure 3 A) ISO3382-1 parameters from six studied virtual concert halls, averages of 24 source positions. B) Spectrograms of early part of impulse responses from one source position.

In all six concert halls the 24 sound source positions were associated with 10 second long anechoic symphony music excerpts by Bruckner and Mahler[21]. Six sources were used for violins, three for violas, three for cellos, two for double basses, four for woodwinds, two for French horns, 3 for trumpets, trombones, and tuba, and finally one sound source for timpani.

In total 19 assessors completed the IVP process in one 2-hours long session. They were screened earlier and they all had earlier participated in other IVP studies. Thus, after a brief introduction they elicited and developed a number of attributes. They then selected the two key attributes and completed the ratings with both music twice, i.e., the rehearsal and the final rounds. Finally, samples were also rated according to preference, providing supplementary data for subsequent analysis.

In total 38 attributes were collected and they are presented with definitions in Table 1. The motivation to apply IVP was to gather all possible attributes with which the samples differ. The IVP is perfect for such studies as the assessors all listens to different aspects of sound, possible different instrument groups and when over 15 assessors do the ratings an extensive list of attributes is generated.

4.2 Reliability of the assessors

As with all statistical data analysis, sufficient data quality is important. Therefore, assessors (naïve or expert) must be able to discriminate between stimuli where there are discriminable differences and to be able to rate in a structured and repeatable fashion for each attribute. Otherwise, assessor are only yielding noise to the dataset. When considering CVP methods, where a common consensus set of attributes is employed, it is relatively easy to test for the *discrimination* ability and *reliability* of assessors and a number of tools have been developed for this purpose as discussed in [10,11,25,26]. However, when using IVP techniques, i.e. individual attributes, such tools are not applicable, and alternative methods are needed. We have addressed the reliability by an iterative approach, by checking 1) whether assessors can replicate their ratings, and 2) whether the individual ratings are connected somehow to the ratings by other assessors.

The IVP process includes practice rating and a final rating, see Fig. 2. Therefore, it is possible to check possible correlation between ratings. As ratings with one attribute are done with all signals, the correlation of two matrices can be done, e.g., with the RV coefficient with the Pearson type III approximation[27]. In addition, the implementation of RV coefficient in `FactoMineR (http://factominer.free.fr/)` returns a p value telling if the correlation is significant or not. For the example data, the correlations of all 38 individual attributes are presented in Fig. 4. It can be seen that only 18 (blue) out of 38 have $p < 0.05$, meaning that they were consistently repeated. Such a low number indicates that the dif-

Table 1 All 38 elicited attributes with their definitions (translated from Finnish to English).

Group	Xnn	Attribute	Low anchor	High anchor	Definition
Width	X5	width	narrow	broad	How wide is sound on left-right axis
	X13	width	narrow	broad	How wide is the sound image
	X35	width	narrow	broad	Width of sound field, how well music envelopes
Envelopment	X18	envelopment	pointlike	enveloping	Sound envelopes, seems to come from everywhere
	X21	envelopment	pointlike	enveloping	Feeling of how well sound envelopes, in particular at mid frequencies
Openness	X33	openness	consice	open	Naturally open, feeling of a space
	X17	openness	obstructed	open	Sound is open, when it sounds "easy" and is not attenuated
	X25	definition	tubelike	defined	Not only mono/stereo. Open is also defined, not foggy
	X7	distance	distant	close	How far the orchestra is
	X3	distance	distant	close	How distant the sound source is
	X28	distance	far	close	Feels like sitting even in front or back row
	X11	richness	less tones	more tones	How many different tones are heard
	X16	nuances	no nuances	many nuances	Are nuances large or boring
	X27	muddy	muddy	clear	Low frequencies muddy, no difference at high freqs
	X6	reveberance	less reverb	more reverb	How much the space reverbs
Bassiness	X24	bassiness	no bass	a lot of bass	Timpani, bassiness
	X30	bassiness	no bass	a lot of bass	When more bass, the sound is more sharp
	X14	bass dominance	less hollow	more hollow	Does bass dominate
	X22	amount of bass	less bass	pronounced bass	Amount of emphasized low frequencies, in particular very low frequencies
Clarity	X12	clarity	muddy	clear	How well the fast passages are clearly heard
	X10	clarity	less warm/clear	more warm/clear	Sound comes as behind the wall, not clear
	X26	diverse	no tones	a lot of tones	Contains lot of tones, harmonics are heard, bass is not muddy
	X23	width	narrow	broad	Broadening of brass and violins, humming of timpani
	X8	envelopment	frontal	enveloping	How well the sound envelopes the listener
	X36	bassiness	no bass	a lot of bass	How basses are dominating the spectrum
	X20	thickness	thin	thick	Size and depth (also color) of sound
	X40	fullness	not full	full	How full and rich the music is
	X37	openness	stuffy	free	Movement of sound in a space
	X34	distance	distant	close	Closeness of sound, fullness
	X39	distance	far	close	At what distance the music comes
Not reliable attributes	X15	sharpness	not sharp	very sharp	Do I hear the melody (sharp) and how instruments blend with each other
	X4	dryness	less dry	more dry	How strongly sound keeps its energy
	X19	articulation	muddy	clear	Definition and clarity of sound, reverb affects
	X29	reveberance	less reverb	more reverb	Reverberance of sound
	X31	reveberance	less reverb	more reverb	Amount of perceived reverb
	X32	bass level	low bass	high bass	Amount of emphasized low frequencies
	X9	muddy	diminishing	reverberant	Can not separate sounds and dry
	X38	muddiness	muddy	clear	How clearly sounds are distinguished from each other

ferences between samples were rather small and possible reasons are 1) the assessors could not rate the samples reliably, 2) they have changed their interpretation, 3) there were not enough training. In the case 1) the final data is noise, but cases 2) and 3) can still be valuable data for the final data analysis[2].

The grouping of IVP attributes is usually done by computing the Euclidean distances (similarity) between attributes. A good tool for such analysis is hierarchical agglomerative cluster analysis using Ward's clustering linkage method. First, all 38 attributes are algorithmically clustered and the result is seen in Fig. 5 (top). One cluster with nine attributes differs the most from the rest of the data. Only one of them (X10) is reliable repeated, thus it might be concluded that the other eight are more or less noise. The two clusters in the lower part of Fig. 5 shows 30 and 18 attributes in each. Comparing these

[2]Ideally, between 4 and 6 repetitions can be needed to firmly establish the reliability of assessors, as discussed by Bech[28], and 2 repetitions, as employed in this study provide a fast initial screening of assessor performance.

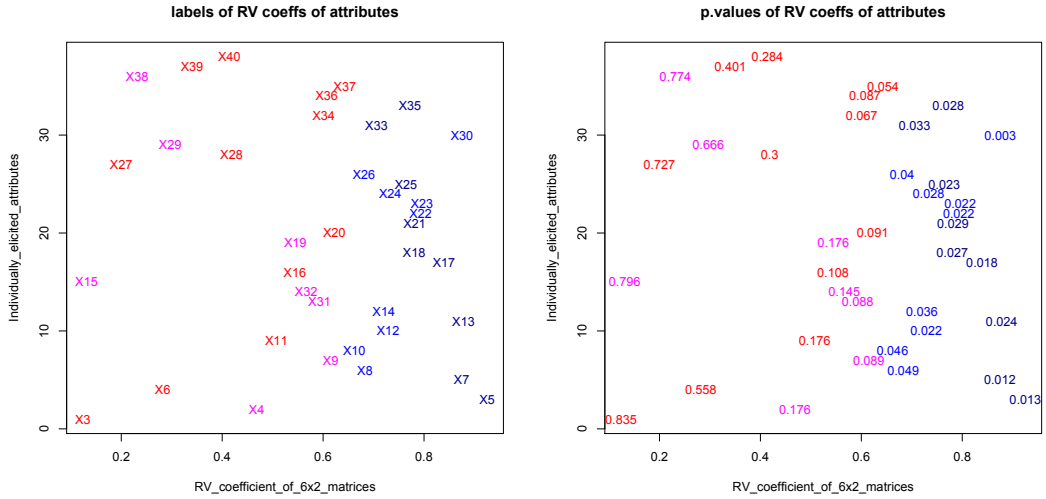

Figure 4 RV coefficients and their p values per attribute between rehearsal and final ratings.

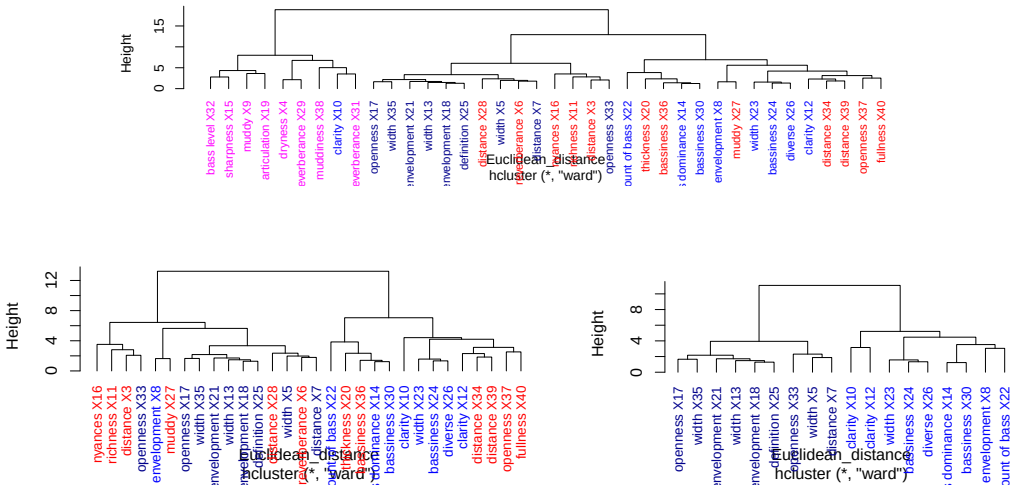

Figure 5 Hierarchical agglomerative cluster analysis. Top: all attributes, Bottom left: two main clusters, Bottom right: only reliable repeated attributes.

two clusters reveals that only one blue attribute (X8) has moved its branch if the unreliably repeated attributes (reds) are included in the analysis. In addition, when the 8 attributes, considered as noise, are removed the clustering still finds only two main clusters, indicating the quality of this attribute data. In other words, the clustering is heuristically monitored in different conditions of discarding suspected unreliable attributes. Attributes for which the clustering is volatile are potentially unreliable.

4.3 Analysis of the IVP data

When the noisy attributes are removed the final rating data can be analyzed with several methods. The purpose of analysis is to order multivariate objects, i.e., samples so that similar objects are near each other and dissimilar objects are farther from each other. Multiple Factor Analysis (MFA)[29,30] is often applied since it derives an integrated picture of the observations and of the relationships between the

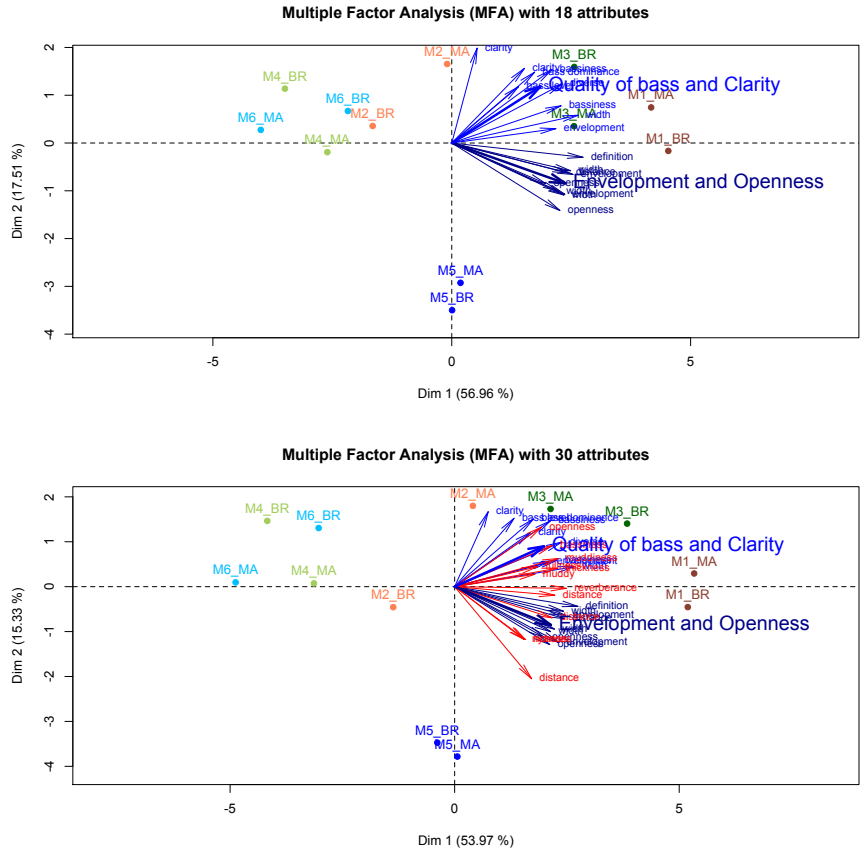

Figure 6 The results of the MFA analyses with 18 and 30 attributes. The biplots visualize both the attributes and samples in the same factorial space.

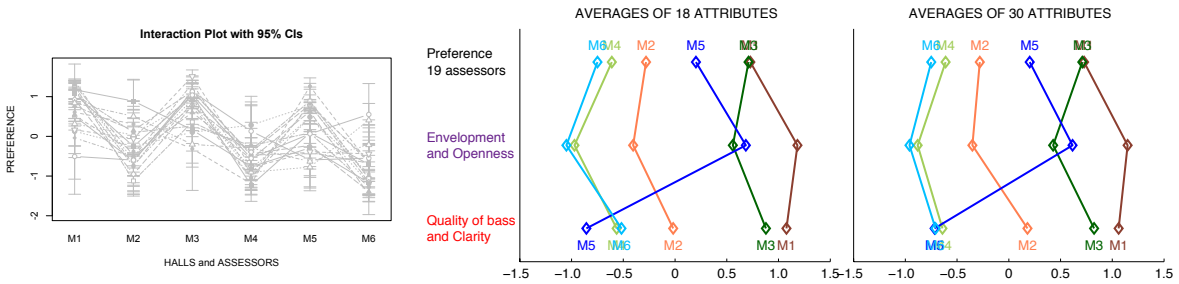

Figure 7 Left: Preference by 19 assessors. Right: Sensory profiles with 30 and 18 attributes.

descriptive attributes. Fig. 6 shows the analysis results for both data sets, with 18 and 30 attributes. The samples are ordinated quite similarly in the plane formed by two principal components. In the case of 18 attributes, the first two principal components explain 74.5% of the variance and with 30 attributes 69.3% is explained. Fig. 6 also illustrates the directions of the largest variance of individual attributes as well as the average perceptual dimension, which is obtained by averaging all attributes that form one cluster (See Fig. 5).

Based on the perceptual dimensions, the sensory profiles for each sample can be formed. Traditionally, such profiles are visualized with spider plots (see Fig. 1), but in our opinion, more informative

illustration is plotted in Fig. 7. The average value for each sample (also average of both signals) shows intuitively the order of samples within the main perceptual characteristics. In addition, the individual preference data of all 19 assessors and average of preference is shown. As there is hardly any difference between the final result with 30 and 18 attributes, we can safely assume that the assessors had insufficient training and they learned to rate the samples during practice prior to actual rating.

The sensory profiles in Fig. 7 illustrate well the IVP study. The shown plots are also the final result of the study with is more deeply motivated earlier[24]. First, the preference ratings are dominated by the differences in Envelopment and Openness. Second, totally diffuse early reflections (M5 and M6) render the sound muddy and weak bass. Third, the median plane early reflections render less enveloping and open sound as expected, but they also deteriorate the quality of bass and clarity, compared to lateral early reflections. Fourth, even the preference ratings did not make difference between M1 and M3, M1 contributes to more enveloping and open sound with slightly clearer sound.

5 CONCLUSIONS

This study describes the successful application of Individual Vocabulary Profiling (IVP) to the characterization of the salient perceptual characteristics of 6 concert hall acoustics with 2 music samples. Stimuli were created using a novel resynthesized orchestra technique, allowing for direct comparison of each hall reproduced in multichannel listening room conditions. 19 selected assessor developed individual attribute sets, of which they selected the 2 most salient for rating. The RV coefficient was employed as a means of evaluating assessor performance and eliminating noisy data. A Multiple Factor Analysis (MFA) was then performed on the individual datasets in order to establish a common perceptual space. Finally, sensory profiles of studied concert halls were used to illustrate the salient perceptual characteristics of this study.

Acknowledgments: The research leading to these results has received funding from the Academy of Finland, project nos. [218238 and 140786] and the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [203636].

6 REFERENCES

1. G. Lorho. *Perceived Quality Evaluation: An Application to Sound Reproduction over Headphones*. PhD thesis, Aalto University School of Science and Technology, 2010. http://lib.tkk.fi/Diss/2010/isbn9789526031965/.
2. H.T. Lawless and H. Heymann. *Sensory evaluation of food: principles and practices*. Aspen Publishers, New York, NY, USA, 1999. 827 pages.
3. S. Bech and N. Zacharov. *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley and Sons Ltd, Chichester, England, 2006.
4. S. V. Legarth, C. S. Simonsen, L. Bramsløw, G. Le Ray, and N. Zacharov. Sensory evaluation of hearing aid performance based on normal-hearing listeners. In *Proc. of the 3rd Int. Workshop on Perceptual Quality of Systems*, Dresden, Germany, September 6-8 2010.
5. N. Zacharov, J. Ramsgaard, G. Le Ray, and C. V. Jørgensen. The multidimensional characterization of active noise cancellation headphone perception. In *Proceedings of the Quality of Multimedia Experience Conference*, Trondheim, Norway, June 21-23 2010.
6. G. Lorho. Individual vocabulary profiling of spatial enhancement system for stereo headphone reproduction. In *the 119th Audio Engineering Society Convention*, New York, NY, USA, October 7-10 2005.
7. G. Lorho, S. V. Legarth, and N. Zacharov. Perceptual validation of binaural recordings for mobile multimedia loudspeaker evaluations. In *AES 38th Int. Conf. on Sound Quality Evaluation*, Piteå, Sweden, June 13-15 2010.

8. T. Lokki, H. Vertanen, A. Kuusinen, J. Pätynen, and S. Tervo. Auditorium acoustics assessment with sensory evaluation methods. In *the International Symposium on Room Acoustics (ISRA2010)*, Melbourne, Australia, August 29-31 2010.
9. ISO 8586-2. *Sensory analysis -- General guidance for the selection, training and monitoring of assessors -- Part 2: Experts*. International Organization for Standards, 1994.
10. N. Zacharov and G. Lorho. What are the requirements of a listening panel for evaluating spatial audio quality? In *Proceedings of the Spatial Audio and Sensory Evaluation Techniques workshop*, University of Surrey, UK, April 6-7 2006.
11. G. Lorho, G. Le Ray, and N. Zacharov. eGauge - a measure of assessor expertise in audio quality evaluations. In *AES 38th Int. Conf. on Sound Quality Evaluation*, Piteå, Sweden, June 13-15 2010.
12. V.-V. Mattila and N. Zacharov. Generalized listener selection (GLS) procedure. In *the 110th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands, May 2001.
13. F. Wickelmaier and S. Choisel. Selecting participants for listening tests of multichannel reproduced sound. In *the 118th Convention of the Audio Engineering Society*, May 2005.
14. S. V. Legarth and N. Zacharov. Assessor selection process for multisensory applications. In *126th Audio Engineering Society Convention*, Munich, Germany, May 7-10 2009.
15. A. Kuusinen, H. Vertanen, and T. Lokki. Assessor selection and behavior in individual vocabulary profiling of concert hall acoustics. In *AES 38th Int. Conf. on Sound Quality Evaluation*, pages 181--190, Piteå, Sweden, June 13-15 2010.
16. R. Kürer, G. Plenge, and H. Wilkens. Correct spatial sound perception rendered by a special 2-channel recording method. In *the 37th Audio Engineering Society Convention*, 1969.
17. M.R. Schroeder, G. Gottlob, and K.F. Siebrasse. Comparative study of european concert halls: Correlation of subjective preference with geometric and acoustics parameters. *Journal of the Acoustical Society of America*, 56(4):1195--1201, October 1974.
18. J. Pätynen, S. Tervo, and T. Lokki. A loudspeaker orchestra for concert hall studies. In *The 7th Int. Conf. On Auditorium Acoustics*, pages 45--52, Oslo, Norway, October 3-5 2008. Institute of Acoustics. Also published in *Acoustics Bulletin* 2009, 34(6), pp. 32-37.
19. J. Merimaa and V. Pulkki. Spatial impulse response rendering I: Analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115--1127, 2005.
20. V. Pulkki and J. Merimaa. Spatial impulse response rendering II: Reproduction of diffuse sound and listening tests. *Journal of the Audio Engineering Society*, 54(1):3--20, 2006.
21. J. Pätynen, V. Pulkki, and T. Lokki. Anechoic recording system for symphony orchestra. *Acta Acustica united with Acustica*, 94(6):856--865, November/December 2008.
22. G. Spikofski and M. Fruhmann. Optimization of binaural room scanning (BRS): Considering inter-individual HRTF-characteristics. In *19th AES Int. Conf.: Surround Sound - Techniques, Technology, and Perception*, June 2001.
23. A. R. Algazi, R. O. Duda, and D. M. Thompson. Motion-tracked binaural sound. *Journal of the Audio Engineering Society*, 52(11):1142--1156, 2004.
24. T. Lokki, J. Pätynen, S. Tervo, S. Siltanen, and L. Savioja. Engaging concert hall acoustics is made up of temporal envelope preserving reflections. *Journal of the Acoustical Society of America Electronic Letters*, 129(5), 2011. In Press.
25. P Schlich. GRAPES: A method and SAS program for graphical representation of assessor performance. *Journal of Sensory Science*, 9:157--169, 1994.
26. P M Brockhoff. Assessor modeling. *Food Quality and Preference*, 9(3):87--89, 1998.
27. J. Josse, J. Pagès, and F. Husson. Testing the significance of the RV coefficient. *Computational Statistics and Data Analysis*, 53:82--91, 2008.
28. S Bech. Selection and training of subjects for listening tests on sound-reproducing equipment. *J. Audio Eng. Soc*, 40(7/8):590--610, 1992.
29. B. Escofier and J. Pagès. Multiple factor analysis. *Computational Statistics & Data Analysis*, 18(1):121--140, 1990.
30. H. Abdi and D. Valentin. Multiple factor analysis. In N.J. Salkind, editor, *Encyclopedia of Measurement and Statistics*, pages 657--663. Sage Publications Ltd., London, UK, 2007.