Audio Engineering Society

Convention Paper 9701

Presented at the 142nd Convention
2017 May 20–23, Berlin, Germany

Combining preference ratings with sensory profiling for the comparison of audio reproduction systems

Tim Walton[1,2], Michael Evans[2], Frank Melchior[2], and David Kirk[3]

[1]*Open Lab, Newcastle University, Newcastle upon Tyne, United Kingdom*
[2]*BBC Research and Development, Salford, United Kingdom*
[3]*Northumbria University, Newcastle upon Tyne, United Kingdom*

Correspondence should be addressed to Tim Walton (`t.walton3@ncl.ac.uk`)

ABSTRACT

One aim of perceptual audio evaluation is to understand the relationships between individual sensory attributes and overall quality of experience. This paper discusses one perceptual evaluation method by which this can be realised. Open Profiling of Quality (OPQ), a method first introduced in the field of visual and audiovisual evaluation, involves psychoperceptual evaluation, sensory profiling and external preference mapping stages and is suitable for use with naïve listeners. Here, a methodological case study is presented in which we discuss the implementation of this method and its adaptation for the comparison of audio reproduction systems.

1 Introduction

A fundamental process in the advancement of audio technology is perceptual evaluation. It is crucial to understand how listeners perceive new technology in order to drive future developments. Objective quality models that predict listener experience are ultimately desirable due to reliability, repeatability and lower resource requirements, however, the complex, multidimensional nature of quality means that their development depends on first fully understanding the relationship between sensory percepts and overall experience. Perceptual evaluation, often by means of listening tests, can help develop insight into this.

Traditionally, standardised listening test methodologies for audio quality evaluation have been focussed on a mean opinion score (MOS) mindset; they use global measures, such as ‘basic audio quality’, to rate a systems acceptability, often in relation to a high quality reference (for example see [1][2]). Such tests are evidently very useful for certain applications (e.g. codec evaluations), however, for the evaluation of innovative technologies that have the potential of providing new experiences to the user (e.g. object-based and immersive audio), a different approach needs to be taken. Reference based methods are not so suitable for evaluating technology that delivers innovative experiences as a high quality reference is generally not available. Furthermore, it could be argued that a more multidimensional Quality of Experience (QoE) mindset is more appropriate for evaluating such technologies. QoE can be defined as “ *the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user’s personality and current state*” [3]. QoE therefore encompasses factors such as context of use, user characteristics and the multidimensional nature of quality.

One goal of perceptual evaluation is to determine the relationships between listener preference and the attributes that play a role in the formation of preference for the relevant content and contexts. By investigating this, insight can be gained in how best to improve the quality of experience for next generation audio technology; if it is known which attributes influence overall preference the most, developers of new technology can utilise this to design for a high quality of experience.

An important factor to take into consideration in perceptual evaluation of audio technology is the listening experience of participants. It cannot be assumed that naïve listeners and experienced listeners will make similar preference and attribute ratings, as shown in [4]. Whereas naïve participants are expected to give holistic, unbiased ratings due to limited knowledge about the stimuli and technology under evaluation [5], experienced assessors are more acute to domain-specific aspects of stimuli, such as audio artefacts. It is therefore desirable to develop test methodologies that are suitable for naïve listeners, as well as experienced listeners.

In this paper, a perceptual evaluation method is discussed that aims to relate preference judgments with relevant attributes. Additionally, it is suitable for naïve listeners. The method - Open Profiling of Quality (OPQ) - was originally introduced by Strohmeier et al. [6] and was discussed in the context of perceptual assessment of next generation audio systems by Walton et al. [7]. Here, a more in depth discussion of the method is given in relation to the experiment presented in [7]. The method is reviewed in terms of its value and applicability for the comparison of audio reproduction systems, including considerations on challenges in the application of the method and possible improvements.

The remainder of the paper is structured as follows: Section 2 introduces the background of OPQ along with a discussion of other methods that combine preference ratings with sensory profiling, a detailed description of the method is given in Section 3 and a discussion of the modifications made to the method, as well as advantages and limitations are presented in Section 4.

2 Related work

Open Profiling of Quality (OPQ) is a method to understand the multidimensional quality of experience, suitable for naïve participants [6]. It is a mixed methods approach meaning that both quantitative and qualitative data are collected leading to a rich understanding of the technology under study. It consists of three primary sections; a psychoperceptual evaluation stage aims to evaluate the degree of overall quality, a sensory profiling stage aims to explore the profiles of the overall quality by means of individual vocabulary elicitation and attribute rating, and finally an external preference mapping stage aims to study the relationship between the overall quality and the quality profiles.

The method was originally developed for the quality evaluation of visual and audiovisual systems, specifically mobile 3D television and video, and has been used in both laboratory and field situations [8][9][10]. Several parallels can be drawn between the original target platforms for OPQ assessment and audio reproduction systems. First and foremost, both visual technology and audio technology reproduce stimuli that are typically heterogeneous and multidimensional in character. Furthermore, both forms of technology have the potential to deliver novel experiences to the user, whether it be through 3D video or binaural audio, and both are used in a wide range of contexts. As OPQ was developed with these aspects in mind, the method is well suited to audio only evaluation as well as visual and audiovisual evaluation.

As such, OPQ has recently been applied by several researchers in the audio community. Walton et al. [7] used an adapted version of the method to compare audio reproduction methods, namely soundbar systems with discrete 5 channel and stereo systems, Sloma [11] applied OPQ for a perceptual comparison of standardised and non-standardised listening rooms and Nowak et al. [12] used OPQ for the assessment of spherical microphone array auralizations.

Combining preference ratings with sensory profiling is not unique to Open Profiling of Quality. Francombe et al. [13] recently used a similar method, which included paired comparisons suitable for naïve participants, to compare spatial audio reproduction systems, although this method has a few notable differences to OPQ. Firstly, whereas OPQ is based on the analysis of individually elicited attributes, the method by Francombe et al. includes a group discussion stage whereby a consensus vocabulary is produced. Furthermore, no explicit attribute rating stage was conducted. Instead a metric called ‘Attribute Score’ was developed, which quantifies the importance of each attribute by considering the frequency with which it was used as well as the size of the preference judgments alongside which it was used.

Zacharov and Koivuniemi [14] introduced a method called Audio Descriptive Analysis and Mapping (ADAM) in the context of perceptual evaluation of spatial audio systems, ranging from mono to eight channel reproduction. In this method a preference rating task is completed by naïve participants in a paired comparison format, a language development task is performed with trained participants, a discussion phase creates a common descriptive language which is then used in an attribute rating stage by trained participants. Finally, partial least squares regression is used to map the subjective preference ratings to the attributes.

Choisel and Wickelmaier [15] conducted an experiment with naïve participants with the aim of quantifying the auditory attributes that underlie listener preference for reproduced multichannel sound. They collected preference ratings via paired comparison judgments and utilised attributes elicited in a previous study to develop ratio scales from probabilistic choice models. Principal components derived from the quantified attributes were then used to predict overall preference.

Zacharov et al. recently presented an interesting method for the assessment of next generation audio systems called the Multiple Stimulus Ideal Profile Method (MS-IPM) [16]. Originally developed in the perfume industry and later applied to hearing aid applications, the method aims to relate overall quality, attribute ratings, and also an ‘ideal profile’. This ideal profile is obtained by asking participants to give an ideal level of each attribute on which the stimuli are being assessed. With regards to attribute elicitation, in this example of the method four specialised expert assessors selected six attributes on which the stimuli were to be rated by the experienced participants.

Studies also exist that relate preference ratings with sensory profiling for perceptual evaluation in other fields of acoustic research. For example, Mattila [17] combined descriptive analysis in the form of paired comparison attribute elicitation and panel discussions, with overall quality judgements for the evaluation of speech quality in mobile communications. In the context of concert hall acoustics, Lokki et al. [18] conducted an individual vocabulary profiling procedure with a triad based elicitation stage and single stimulus attribute rating stage and combined these attribute ratings with preference ratings via preference mapping.

The literature presented above illustrates that there are various examples of combining preference ratings with sensory profiling in the field of audio evaluation. By discussing the OPQ method in this paper, we hope to give experimenters in the audio field another tool to investigate the link between sensory percepts and overall experience for next generation audio technology. The key features for the method discussed here are as follows.

i) It is an individual vocabulary technique, meaning that each participant develops and employs their own attribute list for further rating. An advantage of such methods is that, compared to non-individual methods, participants may be able to better relate to the attributes being used [19].

ii) It is suitable for naïve listeners, as well as experienced listeners.

iii) It is relatively time efficient compared to other similar methods, as the adaption described below only requires two sessions and does not include a panel discussion session.

3 Open Profiling of Quality

In this section, a description of the OPQ method is given. Differences between the original implementation described by Strohmeier et al. [6] and the adapted implementation for the comparison of audio reproduction systems described by Walton et al. [7] are highlighted, although these differences are discussed more thoroughly in Section 4.

3.1 Structure

As previously mentioned, OPQ consists of three primary sections: a psychoperceptual evaluation stage, a sensory profiling stage and an external preference mapping stage, see Figure 1. In the original implementation of the method, the psychoperceptual evaluation and sensory profiling were conducted in different sessions. This was modified in [7], as shown in Figure 2. The purpose of this restructuring was to reduce the duration of the experiment and to aid in the elicitation of attributes that led to listener preference, as discussed in the following section.

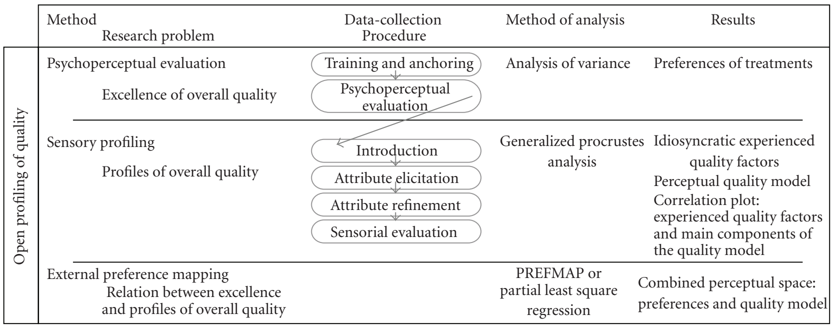

Method
Research problem
Data-collection
Procedure
Method of analysis
Results
Open profiling of quality
Psychoperceptual evaluation
Excellence of overall quality
Training and anchoring
Psychoperceptual evaluation
Analysis of variance
Preferences of treatments
Sensory profiling
Profiles of overall quality
Introduction
Attribute elicitation
Attribute refinement
Sensorial evaluation
Generalized procrustes analysis
Idiosyncratic experienced quality factors
Perceptual quality model
Correlation plot: experienced quality factors and main components of the quality model
External preference mapping
Relation between excellence and profiles of overall quality
PREFMAP or partial least square regression
Combined perceptual space: preferences and quality model

Fig. 1: Overview of the original OPQ structure [6].

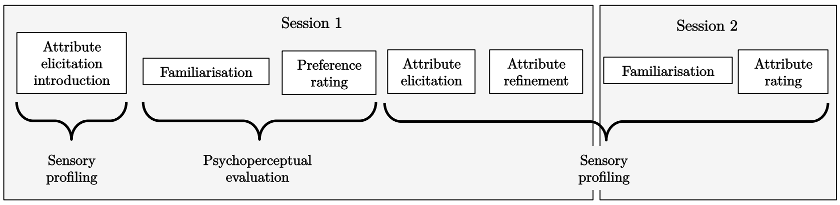


Fig. 2: Overview of the adapted OPQ structure [7]. Session 1 had a total duration of 90-120 minutes and session 2 had a total duration of 60-90 minutes. These took part on separate days.

3.2 Attribute Elicitation Introduction

As this methodology is designed to be suitable for naïve participants, it is necessary to introduce participants to the concept of attribute elicitation. The general idea of this it to ask participants to explain similarities and differences between two stimuli. For example, visual stimuli could be used, such as images [14], or imaginary stimuli, such as two random apples [6]. It is important not to use audio stimuli so as not to bias participants. To conclude this stage the experimenter should relate this mindset of finding similarities and differences back to attribute elicitation for audio stimuli.

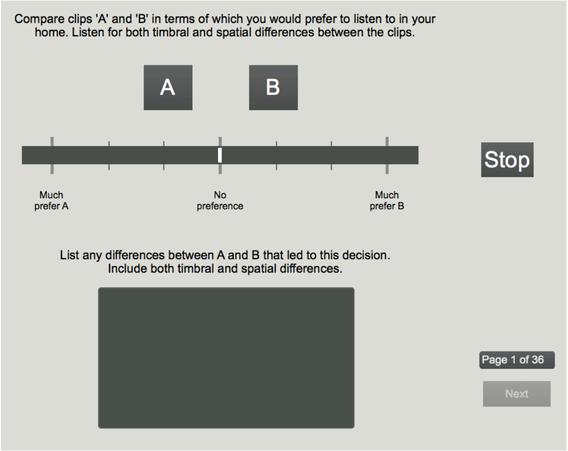
3.3 Familiarisation

Before beginning the rating stages it is important to familiarise participants with both the stimuli that will appear in the experiment and the user interface. By familiarising participants with the scope and range of stimuli that will be used, participants will be able to use the rating scales more effectively which will in turn help reduce scale related bias [20]. A simple method of achieving this is to allow participants to select and play a number of samples that span the range of qualities to be evaluated.

3.4 Preference Rating & Attribute Elicitation

In the original implementation of the method Absolute Category Rating (ACR) with single stimuli was employed for the psychoperceptual evaluation stage. This was adapted to a paired comparison, preference rating method in [7]. It is important to note that ACR is used for the rating of overall quality, which is different to 'preference'. When comparing spatial sound systems without introduced degradations, it could be the case that quality is perceived as equally high for all systems, even though participants may have certain preferences. For this reason, preference ratings were deemed to be more suitable than overall quality ratings for the case of comparing reproduction systems. Advantages of paired comparisons is that they are powerful when comparing systems with small differences whilst still being simple for untrained participants, although this is at a cost of an increased experiment duration. The following instructions were given to participants: "Compare clips 'A' and 'B' in terms of which you would prefer to listen to in your home. Listen for both timbral and spatial differences between the clips". A continuous slider was used for the rating, with offset end anchors so as to reduce end-of-scale effects [21].

Whereas attribute elicitation was conducted in a separate session to preference rating in the original implementation of the method, in [7] attribute elicitation was carried out simultaneously with preference rating, as seen in Figure 3. The aim of this was to encourage participants to list attributes that were directly related to their given preference ratings. Moreover, with this format there were two sections that involved listening instead of three with the original format, possibly reducing listener fatigue. Participants were instructed to "List any differences between A and B that led to this decision. Include both timbral and spatial differences." Elicitation from paired comparisons was seen to be effective for use with naïve participants. Paired and triadic comparisons have previously been used for attribute elicitation in other methods, such as the Repertory Grid Technique [22].



Compare clips 'A' and 'B' in terms of which you would prefer to listen to in your home. Listen for both timbral and spatial differences between the clips.

List any differences between A and B that led to this decision. Include both timbral and spatial differences.

Fig. 3: Graphical user interface for preference rating and attribute elicitation stage. In [7], each page was a comparison of two reproduction systems for the same content item.

3.5 Attribute Refinement

For accurate profiling it is necessary to refine the list of attributes that participants develop. In the adaption in [7] this takes place directly after the rating stage. Participants are presented with all of their elicited attributes and are asked to refine them with help from the experimenter. This includes:

i) Selecting unique attributes, i.e. those that do not cover the same aspect of quality, including grouping opposite terms together so that either the positive or negative remain.

ii) Asking participants to describe the remaining attributes in their own words, so that the experimenter can be sure that the participants fully understand which aspect of quality is being described. To ensure that the attributes are unique, participants are asked to explain the difference between attributes that sound similar.

iii) Reducing the number of attributes to a suitable number for the attribute rating stage. Participants are asked to select the '*n*' most important attributes that influenced their preference ratings throughout the test. In [7], 8 was chosen as an appropriate number although this can be changed depending on the study. For some participants who have not generated more than *n* attributes, this stage is not necessary.

3.6 Attribute Rating

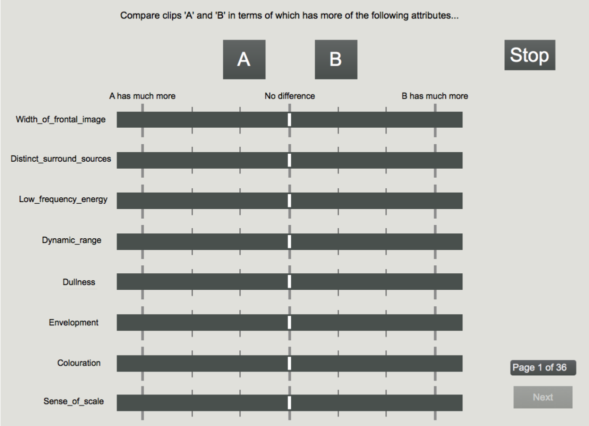
The aim of the attribute rating stage is to quantify the strength of the developed attributes for each stimulus. As with the preference rating stage, a paired comparison method is used in [7]. Participants are asked to rate which stimulus, A or B, has more of the listed attributes, as seen in Figure 4. This is in contrast to a single stimulus method described in [6]. Rating all attributes in the same trial for each paired comparison has the advantage of reducing the test duration compared to rating attributes in succession, however, it should be noted that one disadvantage could be inter-attribute correlations, i.e. a general preference for a stimulus might show as increased attribute ratings in favour of that stimulus.

3.7 Analysis

This section gives a brief overview of the methods used when analysing results from OPQ. Specifically, methods are presented suitable for evaluating a paired comparison implementation of OPQ, as described in the previous section.

3.7.1 Participant consistency

A measure of participant consistency when using paired comparison ratings is circular error rates [23]. A circular error occurs when a participant makes an inconsistent judgment on a triad of stimuli. For example, a circular error would occur if a participant preferred stimulus A to B, preferred stimulus B to C, but preferred



Compare clips 'A' and 'B' in terms of which has more of the following attributes...

Fig. 4: Graphical user interface for attribute rating stage with example attributes. Attributes correspond to each participant's individually elicited attributes. In [7], each page was a comparison of two reproduction systems for the same content item.

stimulus C to A. It is possible to calculate a circular error rate in percent for each participant by comparing the number of circular errors associated with each participant and the maximum possible number of circular errors. A high circular error rate indicates that the participant was not paying attention, that they altered their assessment criteria as the test progressed, or that they found the test challenging. In [7] only the preference ratings were used for participant consistency analysis, however, it is also possible to use attribute ratings.

3.7.2 Preference ratings

Before further analysis can take place, raw data from the paired comparison preference ratings need to be converted to preference scores. This is done through the following steps.

i) If the continuous preference rating scale was not setup to give values between ±1, the preference ratings for each paired comparison need to be scaled to lie in the range of ±1, where -1 corresponds to full preference for stimulus ‘A’ and +1 corresponds to full preference for stimulus ‘B’.

ii) If P_{ij} is the preference probability of stimulus i versus stimulus j, it is assumed that

$$P_{ij} = -P_{ji}. \tag{1}$$

That is, a negative probability of preference P_{ij} means stimulus j is preferred to stimulus i. From these preference probabilities, preference scores can be calculated with

$$S_i = \sum_{j \neq i} P_{ij}, \tag{2}$$

where S_i is the preference score for stimulus i.

iii) If ‘n’ reproduction systems are being compared, the above preference scores can have values in the range of $\pm(n-1)$. These scores are then scaled to lie in the range of ±1 so that +1 corresponds to full preference towards a reproduction system.

This processed preference score data is then used for further analysis, such as analysis of variance (ANOVA).

3.7.3 Sensory profiling

As the attribute ratings are also made through paired comparisons the raw paired attribute ratings are converted to attribute scores using the same method as with the preference scores. For each participant this results in an M x N matrix (or configuration) of attribute ratings, where M is the number of test items and N is the number of individual attributes. The individual participant matrices are then concatenated to form a complete attribute matrix of items x attributes.

This dataset of attribute ratings is then processed with multidimensional data analysis methods. Generalised Procrustes Analysis (GPA) is used to reduce scale effects and to obtain a consensus configuration. This is achieved by rotating and transforming the configurations by minimising the residual distance between the configurations and their consensus. Secondly, Principal Component Analysis (PCA) is used on the dataset from the GPA procedure.

By identifying the ‘elbow point’ in the cumulative variance data from the PCA analysis, it can be decided which components should be used to form the perceptual space. Components that appear before the elbow are retained for further analysis [24]. For example, in [7], the first two components described 88% of the variance whereas the first three described 92%, so it was justified to form a perceptual space out of the first two components only.

Interpretation of the sensory profiling data is achieved by plotting both the individual attributes and the test

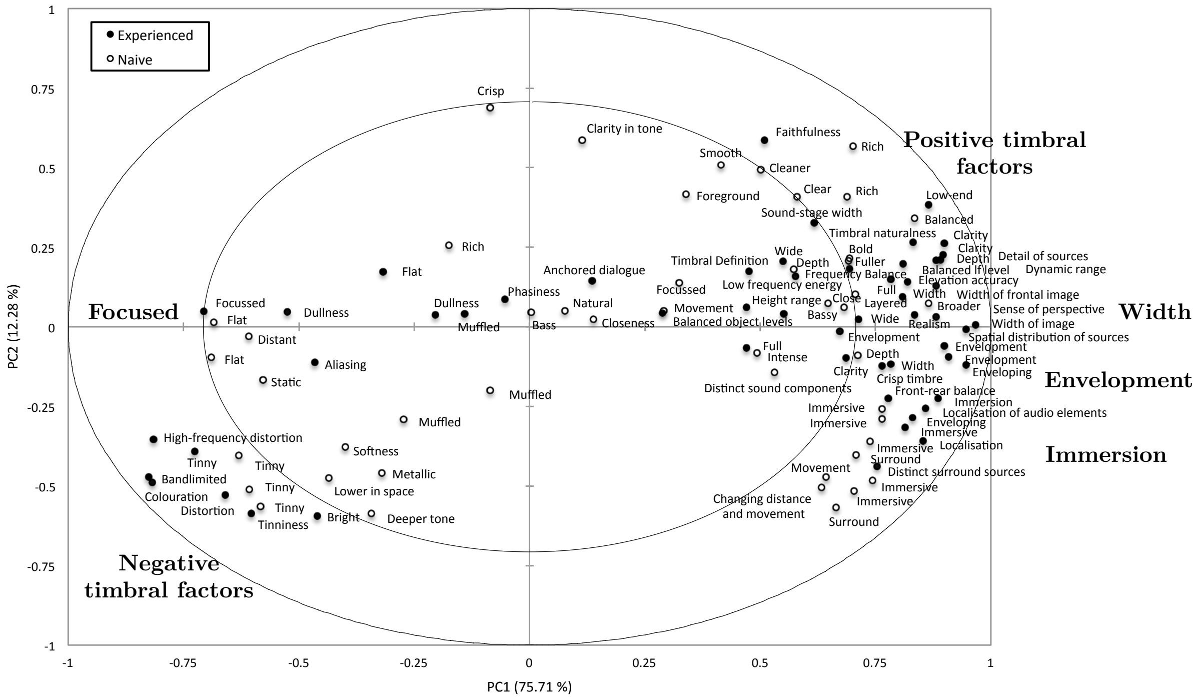


Fig. 5: Example PCA correlation loadings with attributes in the space of PC1 and PC2. The inner and outer circles represent 50% and 100% explained variance respectively. [7]

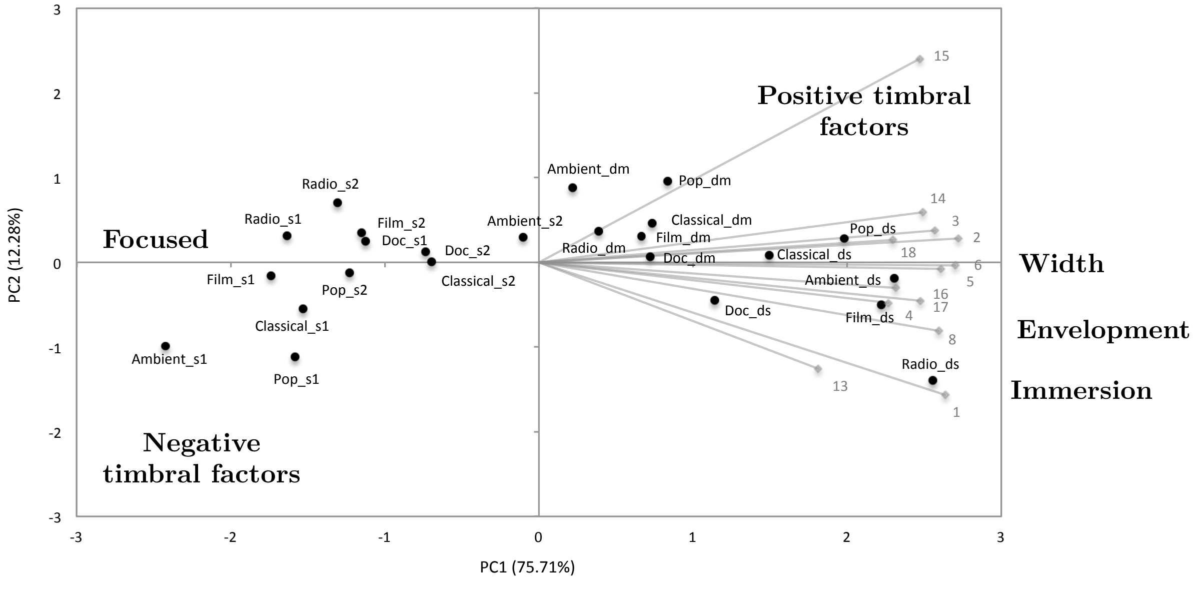


Fig. 6: Example objects and participants' preferences in the preference map of PC1 and PC2. 'ds' = discrete surround, 'dm' = stereo downmix, 's1/2' = soundbar 1/2. [7]

stimuli in the perceptual space, figures 5 and 6 respectively. Attributes that are located towards the edge of the perceptual space describe more variance than those that are located towards the centre. By examining the location of the attributes it is possible to assess if participants used similar sounding attributes in similar ways. Attribute clusters may be noticeable and these should be identified by the examiners. This can be achieved either through visual inspection or a more formal cluster analysis. For example in Figure 5, identified attribute clusters include width, envelopment, immersion and positive and negative timbral attributes. From the identified attribute clusters it may be possible to label the dimensions of the perceptual space. In Figure 5, it is seen that the dimension PC1 predominantly corresponds to spatial factors ranging from 'focussed' at negative values to 'width' at positive values. PC2, however, is not so easily identified and this highlights the fact that the labelling of dimensions is not always a trivial task. The attribute clusters relating to positive and negative timbral effects are found on the diagonal of the perceptual space. This suggests that for the stimuli used, timbral and spatial factors were correlated and non-orthogonal.

Once groupings are found in the attributes and an understanding of the perceptual dimensions has been obtained, the positions of the test stimuli in the perceptual space can be related to these. For example, in Figure 5 a grouping of negative timbral terms is found in the bottom left of the perceptual space. When looking at the object plot in Figure 6, it is seen that these attributes correspond to the system labelled 's1'. Likewise, system 'ds' corresponds to attributes related to width and envelopment.

External preference mapping is then conducted by combining the preference data with the attribute data. The result of this is a map of participants' preferences on the perceptual space, as seen in Figure 6. By combining both the preference data and sensory profiling data, a rich understanding of the systems under study is gained.

4 Discussion

In the previous section it was seen that with OPQ, preference ratings and sensory profiling data are gathered and analysed in order to develop an understanding of the technology under evaluation. Originally developed with the assessment of visual and audiovisual stimuli in mind, the method transfers well to the assessment of audio only material as, like visual stimuli, audio stimuli are typically heterogenous and multidimensional in character. The method's main distinction between other audio evaluation methods that combine preference rating and sensory profiling is that OPQ is an individual vocabulary technique and does not involve a consensus vocabulary at any stage.

Several modifications to the original implementation of the method were made in the audio comparison use case described, with the aim of making the method more suited to the application. Firstly, quality ratings were modified to preference ratings for the reason that when comparing audio reproduction systems without introduced degradations, it could be the case that quality is perceived as equally high for all systems, even though listeners may have certain preferences. Secondly, a paired comparison approach was taken throughout the modified method in comparison to a single stimulus approach in the original implementation. One advantage of using a paired comparison method is that such an approach is able to provide a high degree of discrimination between stimuli whilst still being suitable for naïve participants. A high degree of discrimination is an important feature when comparing audio reproduction systems with participants who are not necessarily used to critical listening. Additionally, with a paired comparison approach it is possible to make consistency checks by means of circular error rates. Finally, the structure of the method was modified so that the preference rating and attribute elicitation stages occurred simultaneously. This meant that participants were only required to listen to the stimuli in two sessions not three, possibly reducing listener fatigue.

It is also worth discussing limitations of the modifications made above and the OPQ method in general. The number of comparisons to be made in a full factorial paired comparison approach rapidly increase with the number of systems to be assessed. This puts a lower limit on the number of systems than can be assessed with a paired comparison approach compared to a single stimulus approach. A limitation of using naïve participants for sensory profiling is that they are less acute to certain attributes compared to trained listeners. Depending on the purpose of the study, this may or may not be an issue. An alternative approach would be to use naïve participants for the preference rating stage and trained listeners for the sensory profiling stage.

5 Conclusion

Combining preference ratings with sensory profiling is a useful technique to evaluate audio technology that delivers innovative user experiences. By mapping user preferences to individually elicited attributes, it is possible to gain a deep understanding of the formation of preference for the technology, content and context in question. As a result, this enables developers of new technology to understand how to improve the quality of experience for next generation technology. In this paper we have discussed one method by which to achieve this - Open Profiling of Quality. An overview of the method was given in relation to its adaption in a recent study. It is hoped that this method may be of use for other researchers in the field of audio.

Acknowledgements

This work was supported by an ESPRC industrial CASE studentship in partnership with the BBC.

References

[1] ITU-R, "BS.1116-3 Methods for the subjective assessment of small impairments in audio systems," International Telecommunication Union, 2015.

[2] ITU-R, "BS.1534-3 Method for the subjective assessment of intermediate quality levels of coding systems," International Telecommunication Union, 2015.

[3] Qualinet, *Qualinet White Paper on Definitions of Quality of Experience, Version 1.2*, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, 2012.

[4] Rumsey, F., Zielinski, S., Kassier, R., and Bech, S., "Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences," *Journal of the Acoustical Society of America*, 117(6), pp. 3832–3840, 2005.

[5] Jack, F. R. and Piggott, J., "Free choice profiling in consumer research," *Food Quality and Preference*, 3(3), pp. 129 – 134, 1991–1992.

[6] Strohmeier, D., Jumisko-Pyykkö, S., and Kunze, K., "Open Profiling of Quality: A mixed method approach to understanding multimodal quality perception," *Advances in MultiMedia*, 2010, pp. 3:1–3:17, 2010.

[7] Walton, T., Evans, M., Kirk, D., and Melchior, F., "A subjective comparison of discrete surround sound and soundbar technology by using mixed methods," in *Audio Engineering Society Convention 140*, 2016.

[8] Strohmeier, D., Jumisko-Pyykko, S., and Eulenberg, K., "Open Profiling of Quality: Probing the method in the context of use," in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, pp. 7–12, 2011.

[9] Strohmeier, D., Jumisko-pyykko, S., Kunze, K., and Bici, M. O., "The extended-OPQ method for user-centered quality of experience evaluation: a study for mobile 3D video broadcasting over DVB-H," *EURASIP Journal on Image and Video Processing*, 2011, pp. 1–24, 2011.

[10] Strohmeier, D., "Open Profiling of Quality: a mixed methods research approach for audiovisual quality evaluations," *SIGMultimedia Rec.*, 4(4), pp. 5–6, 2012.

[11] Sloma, U., "Evaluation of quality features of spatial audio signals in non-standardized rooms: Two mixed method studies," in *Audio Engineering Society Convention 140*, 2016.

[12] Nowak, J., Jurgeit, K.-P., and Liebetrau, J., "Assessment of spherical microphone array auralizations using Open-Profiling of Quality (OPQ)," in *QoMEX*, 2016.

[13] Francombe, J., Brookes, T., Mason, R., and Woodcock, J., "Determining and labelling the preference dimensions of spatial audio replay," in *QoMEX*, 2016.

[14] Zacharov, N. and Koivuniemi, K., "Unravelling the perception of spatial sound reproduction: Techniques and experimental design," in *Audio Engineering Society Conference: 19th International Conference: Surround Sound - Techniques, Technology, and Perception*, 2001.

[15] Choisel, S. and Wickelmaier, F., "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference," *Journal of the Acoustical Society of America*, 121(1), pp. 388–400, 2007.

[16] Zacharov, N., Pike, C., Melchior, F., and Worch, T., "Next generation audio system assessment using the multiple stimulus ideal profile method," in

2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), 2016, doi:10.1109/QoMEX.2016.7498966.

[17] Mattila, V.-V., "Descriptive analysis of speech quality in mobile communications: Descriptive language development and external preference mapping," in *Audio Engineering Society Convention 111*, 2001.

[18] Lokki, T., Pätynen, J., Kuusinen, A., and Tervo, S., "Disentangling preference ratings of concert hall acoustics using subjective sensory profiles," *Journal of the Acoustical Society of America*, 132(5), pp. 3148–3161, 2012.

[19] Berg, J., "How do we determine the attribute scales and questions that we should ask of subjects when evaluating spatial audio quality?" in *Spatial Audio & Sensory Evaluation Techniques*, Guildford, UK, 2006.

[20] Bech, S. and Zacharov, N., *Perceptual audio evaluation - Theory, method and application*, John Wiley & Sons, Ltd., 2006.

[21] Zielinski, S., Rumsey, F., and Bech, S., "On some biases encountered in modern audio quality listening tests - a review," *J. Audio Eng. Soc*, 56(6), pp. 427–451, 2008.

[22] Choisel, S. and Wickelmaier, F., "Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound," *J. Audio Eng. Soc*, 54(9), pp. 815–826, 2006.

[23] Parizet, E., "Paired comparison listening tests and circular error rates," *Acta Acustica united with Acustica*, 88(4), pp. 594–598, 2002.

[24] Lawless, H. and Heymann, H., *Sensory evaluation of food - Principles and practices*, Springer, 2010.