An Investigation of Qualitative Research Methodology for Perceptual Audio Evaluation

(Licentiate thesis)

By Dan Nyberg
Abstract
This thesis investigates whether a qualitative research method, using phenomenological interviews and analysis, can be successfully applied to perceptual audio research, a field heretofore that has mainly used quantitative methods. The method is investigated by studying the types of information elicited by the method and the information’s usefulness and relevance to the conducted study. The qualitative method is applied in three different conditions: a non-experimental condition, an experimental condition, and an experimental condition using mixed-methods. The thesis also identifies implications associated with using a qualitative method in a quantitative field of research, implications that researchers should acknowledge and consider. All scientific criteria in which the quantitative research is judged cannot directly be applied to a qualitative method. A qualitative method has to be judged on its own framework, departure points, and scientific criteria. The information elicited from the qualitative method contains information that supports known knowledge and adds new knowledge. It supplements the accessibility to the subjects’ perceptions and used methods when conducting a perceptual evaluation task. In conclusion, a qualitative research method that consists of phenomenological interviews and analyses can be successfully applied in all the tested conditions.
Preface

This licentiate thesis contains work performed at Luleå University of Technology, the institution of Art, Communication and Education in Piteå, Sweden, during 2009-2012. The thesis include research from the Vir Music Project financed by the European Union program Interreg IV A Nord, research in cooperation with the Swedish Radio, as well as individual research conducted by the author.

This thesis would not have been accomplished without valuable comments, discussions with supervisors, colleagues and the project and co-operation members. So I would like to take the time to acknowledge these persons.

Special thanks to my supervisor Associated professor Jan Berg for letting me have the opportunity to develop my intellectual capacity and making this thesis a reality and for having the openness to inspire the use of a methodology new to the research field. Many thanks to my co-supervisor Dr Anders Persson, for the rich and valuable comments and discussions on the work, including large and small issues.

Many thanks to all colleagues and members involved in the project and co-operation in which enabling the research. From the Vir Music project the author would like to especially thank Morten Amundsen and Noa Nakai for a successful collaboration on making the Master Classes happen both logistically and with equipment. Thanks to Sture Brändström, Erik Lundström, Outi Nissi, Jenny Kilponen and Marrit Rajamäki for planning the master classes and helping with the design and distribution of questionnaire.

Thank you to Lars Mossberg, Christofer Bustad and Lars Jonsson of the Swedish Radio for letting my research be a part of your research. Special thanks to Lars Mossberg and Christofer Budstad for a successful co-operation with the listening test in Piteå and for conducting the interviews in Stockholm.

The author would also like to thank the PhD student colleagues, Jonas Ekeroot and Oscar Gedda, for valuable discussions and comments on ontological and epistemological considerations. A special thank to Jonas Ekeroot for having the time to give comments on the articles during the writing process. Thanks to Lars Hallberg and Christer Wiklund for giving the financial support. Thanks to all participants in all studies, without your participation this work would not have been possible.

Hopefully the work and the results from this thesis will be “a thing that wouldn’t leave”.

D. Nyberg

Licentiate thesis

2012
**Table of Content**

Abstract ....................................................... 2
Preface .......................................................... 3
1. Introduction .............................................. 5
2. Theory and background ......................... 6
   2.1. Definition of quality .............................. 6
   2.2. Point of departure ................................ 8
   2.3. Distinctions between a qualitative point of
        departure and a quantitative point of
        departure ............................................. 10
   2.4. Perceptual audio research ....................... 10
   2.4.1. Physical domain ................................ 12
   2.4.2. Sensory domain ................................ 12
   2.4.3. Affective domain ............................... 13
   2.4.4. Response format ................................ 13
   2.4.5. Implications of perceptual audio
           research ......................................... 13
   2.5. Mixed methods research ......................... 14
   2.5.1. Implications of mixed methods research ... 15
   2.6. Qualitative methods ............................... 15
   2.6.1. Phenomenology .................................. 16
   2.6.2. Hermeneutics .................................... 16
   2.6.3. Grounded theory ............................... 17
   2.6.4. Comparison of qualitative
           research ........................................... 17
   2.7. Summary of theory and
        background ......................................... 18

3. Research question and focus ............ 18
   3.1. Motivation ........................................... 18
   3.2. Aim ................................................. 19
   3.3. Objectives .......................................... 19

4. Summary of papers ......................... 20
   4.1. Paper 1 ............................................. 20
   4.2. Paper 2 ............................................. 21
   4.3. Paper 3 ............................................. 22

5. Summary of results ......................... 23
   5.1. Types of elicited information .................... 23
   5.2. Relevant information for the study .............. 24
   5.3. Usefulness of the information .................... 24

6. Discussion ........................................... 25
   6.1. Relation to quality aspects ....................... 25
   6.2. Framework of Validity and reliability ............ 25
   6.3. Interviews and analyses ........................... 26
   6.4. Information elicited from the studies ............ 26
   6.5. Using the qualitative methods ..................... 27
   6.6. Original contribution ............................. 28
   6.7. Further work ...................................... 28

References ............................................... 28

**Appended papers:**

Paper 1:
A qualitative approach to evaluation of perceived qualities of audio and video in a distance education context

Paper 2:
Perceived audio quality of realistic FM and DAB+ radio broadcasting systems

Paper 3:
A Descriptive Method for Unravelling the Composition of the Black Box of Affective Judgment in Sound Quality Evaluation
1. Introduction
Perceptual audio evaluation requires research, prediction, evaluation, and generalization of perceived sound quality. These procedures frequently involve using models and numeric data collection. This way of quantitatively evaluating perceived audio quality has been highly effective and the main mode of investigation; such conditions allow the researcher to control for all parameters during an experiment and involve rigorous control over the parameters in the experiment. This strategy is done to collect answers that are related to the factor under investigation and to control and estimate any bias. In addition, this strategy often does not allow test subjects to add information about their own perception of the sound quality.

This type of investigation has several implications. First, as perceived sound quality can be highly subjective, controlling what factors the subject evaluates eliminates the ability of the subject to add information about a perception. This limitation can mean that the researcher will miss important information related to the evaluation. Second, controlling all parameters might not be possible in all situations. This is especially true in real-life situations where evaluation of sound quality is needed, i.e., where interaction (e.g., talking or collaboration) between subjects cannot be interrupted.

To deal with these implications, a new data collecting method and analysis that can add more detailed information and a different aspect on a perception might be valuable. This approach shifts the focus to qualitative methods frequently used in the social sciences, humanities, and arts. These methods focus on the individual subject’s perception and experience. By allowing the subjects to describe and reflect on their perception in their own words (during, after, or even before), a sound quality investigation could be even more useful. The approach of letting subjects use their own words to describe their perception, in the form of attributes, is not new to this field; it has been successfully applied in the field of perceptual audio evaluation in a more quantitative manner [1]. However, this shift to entirely qualitative methods comes with implications. For example, the researcher has to be aware how qualitative methodology differs from perceptual audio evaluation methods.

This study investigates whether a qualitative method, phenomenological interviews and analyses (techniques often used in the humanities, social sciences, and arts), can generate useful and relevant information to the field of perceptual audio evaluation. In this study, the terms audio quality and sound quality are used interchangeably.

The thesis begins with a background and theory section that presents definitions of perceived quality, the ontological and epistemological departure points, a discussion on how qualitative and quantitative research differ, other points of departure, and the research focus. The thesis contains three publications/studies, all of which use a qualitative method. The method is applied in three different conditions to achieve the thesis aim:

- a non-experimental condition;
- an experimental condition; and
- an experimental condition using mix-methods.
2. Theory and background
This section defines the point of departure and what is meant by quality. The foundation in which the research rests on is also presented, incorporating previous research and methodological considerations.

2.1 Definition of quality
To study perceived audio quality, a clarification of the term quality is needed. Martens and Martens present four definitions of quality [2], seen below:

- **Definition 1: Quality as qualitas.**
  Underlying assumption: People *describe* the world. The focus is on the inherent properties of an entity (e.g., a system) so as to describe the properties of an entity objectively.

- **Definition 2: EXCELLENCE/GOODNESS:**
  Underlying assumption: People *evaluate* the world in a cultural context. This definition considers quality as the degree of excellence. This definition is in conflict with definition 1. And seen from a market study point of view, there is a danger of defining a degree of excellence from a single individual. In a market study approach, inter-subjective patterns (i.e., patterns of responses common for a whole group of subjects) are sought. Thus a statistical representative quantity of a specified target group is needed.

- **Definition 3: STANDARDS:**
  Underlying assumptions: People *relate* the descriptions of an entity to an *evaluation* of good and bad, making quality a relational concept. Quality is thus a combination, to a certain extent, of both definitions 1 and 2. The focus is on the description of an entity and the stated/implied requirements.

- **Definition 4: AN EVENT:**
  Underlying assumption: People *implicitly experience* quality as an event that *can only be lived*. The focus is shifted from a third person point of view to a first person view. The quality is seen as a subjectively experienced event, combining knowledge about “what is” with feelings about “what is best”. The form of quality approach can be considered as the ultimate subjective and phenomenological view.

The first three definitions of quality are often seen in perceptual audio research. The focus varies between describing a sensation objectively and evaluating the audio subjectively, depending on the research focus. For example, when defining basic audio quality [3], an assessment of the overall audio quality is sought; in Lorho’s definition [4] of quality, a specific character of quality is under investigation (e.g., spatial impression). As Jekosh [5] points out, it is also important that there is a distinction made between quality seen from a producer’s perspective and quality seen from the user’s perspective: the former relates to product sound quality and the latter relates to perceived sound quality. Table 1 shows several commonly used definitions of sound quality.
Sound quality can also be categorized into four conceptual layers based on the amount of abstraction involved in the methods based on the properties of their source perceptions (Table 2). Auditive Quality uses the lowest level of abstraction, and Aural Communication Quality uses the highest level of abstraction [6].

This research relates quality aspects to perceived sound quality and not the product sound quality. Furthermore, the concept used in this thesis coincides (largely) with definition 3: Quality is a rational concept between evaluation of good and bad and the description of an entity. However, few perceptual audio studies focus on the fourth definition of quality. This inconsistency raises an interesting question: What information can be elicited when investigating perceptual audio quality with a focus on the subjective?

Table 1: Examples of definitions of quality in perceptual audio research.

<table>
<thead>
<tr>
<th>ITU-Recommendations</th>
<th>BS. 1116 [3], definition of Basic Audio Quality (BAQ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lorho’s [4] definition of quality</td>
<td>“A measure of the distance between the character of an entity under study and the character of a target associated with this entity”.</td>
</tr>
<tr>
<td>Jekosh [5] definitions of product and perceived sound quality</td>
<td>Product-sound quality:</td>
</tr>
<tr>
<td></td>
<td>“Product-sound quality is a descriptor of the adequacy of the sound attached to a product. It results from judgments upon the totality of auditory characteristics of the said sound- the judgments being performed with reference to the set of those desired features of the product which are apparent to the user in their cognitive, actual and emotional situation”.</td>
</tr>
<tr>
<td></td>
<td>Sound quality:</td>
</tr>
<tr>
<td></td>
<td>“Result of an assessment of the perceived auditory nature of a sound with respect to its desired nature”.</td>
</tr>
</tbody>
</table>
Table 2: Synopsis of conceptual layers of sound quality (Taken from Blauert and Jekosh [6]).

<table>
<thead>
<tr>
<th>Conceptual Aspects</th>
<th>Examples of Issues</th>
<th>Suitable Measuring Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditive Quality (Classical Psychoacoustics)</td>
<td>Perceptual properties such as loudness, roughness, sharpness, pitch, timbre, and spaciousness</td>
<td>Indirect scaling: thresholds, difference limens, points of subjective equality</td>
</tr>
<tr>
<td>Aural-scene Quality (Perceptual Psychology)</td>
<td>Identification and localization of sounds in a mixture, speech intelligibility, audio perspective incl. distance cues, scenic arrangement, tonal balance, and aural transparency.</td>
<td>Discretic: semantic differential, multi-dimensional scaling. Syncretic: scaling of preference, suitability and/or appropriateness, benchmarking against target sounds</td>
</tr>
<tr>
<td>Acoustic Quality (Physics)</td>
<td>Sound-pressure level, impulse response, transmissions function, reverberation time, sound-source position, lateral-energy fraction, and inter-aural cross correlation</td>
<td>Instrumental measurements with physical equipment for measurement of elasto-dynamic vibrations and waves, including appropriate signal processing</td>
</tr>
<tr>
<td>Aural-communication Quality (Communication Sciences)</td>
<td>Product-sound quality, comprehensibility, usability, content quality, immersion, assignment of meaning, and dialogue quality</td>
<td>Psychological (cognitive) tests, particularly in realistic use cases, e.g., the product in use, the audience in concert, etc., questionnaires, dialogue tests, comprehension test, usability tests, and market surveys</td>
</tr>
</tbody>
</table>

2.2 Point of departure

This section presents some of the crucial ontological and epistemological underpinnings that inform this research. According to Hansson [7], science provides us with common parts of our worldviews. This knowledge should be inter-subjective (common to all humans) and aim to obtain knowledge of how things really are. Thus, science should strive to produce objective knowledge.

The demands of objectivity and inter-subjectivity can be stated in three assumptions on which science is based ([7] p. 11):

1. There is a real world that is independent of our senses.
2. This real world is common to all of us.
3. With combined forces we can achieve, or at least approach, knowledge about this real world that is common to us all.

Hansson [7] states that we have no direct access to physical reality (we only have access to the impressions our senses perceive). There is no simple or direct way to check that a statement is objective although there is a way to check whether a statement is inter-subjective, whether anyone could arrive at the same conclusions using critical testing.
This understanding, according to Hansson, makes inter-subjectivity a requirement for scientific knowledge. One important distinction between objectivity and inter-subjectivity is that inter-subjectivity can be reached without objectivity since people can share prejudices. Objectivity, on the other hand, includes inter-subjectivity, since the demands for objectivity are the same for all people. Thus we try to describe a single reality as accurately as possible [7].

Patton [8] argues that all our understanding comes from our sensory experience of a phenomenon and this experience should be described, explicated, and interpreted. However, this approach is difficult since descriptions of experience and interpretation are highly intertwined. Hence, according to Patton, there is no separate or objective reality for people: there is only what people know their experience is and means, so the subjective experience incorporates the objective thing and people’s reality [8]. For Hansson, a researcher should always strive towards objectivity even if it is not possible to achieve fully [7]. This is essential both for the production of knowledge (facts) and for knowledge that results in action:

*When a historian makes an account of how World War II started, or a chemist explains a particular chemical reaction, we expect their statements to refer to and describe reality; they should not report personal suggestions or ideas.* ([7] p.11)

This view on knowledge is the foundation of this work, where the subjects’ perspectives are the focus. It also illustrates the difficulties with perceptual audio evaluation. Perceptions of audio quality are interpretations made by the listener, but the distinction of what is perceived (objectively) and how the listener interprets this perception are intertwined from the researcher’s perspective. Because trained listeners can consistently discriminate differences/similarities, this problem can be minimized. However, trained listeners can be too discriminating in their evaluation when looking for an example of the overall quality. In these cases, untrained and naïve listeners might be more suitable because they do not analyse the sound to the same degree as trained listeners [4, 9].

Blauert and Jekosch [6], using another point of departure, present a new way to categorize perceptual audio research methods, perceptionism. Perceptionism starts with the assumption that the world is constructed in our brain and is composed of feelings (hunger, pain, etc.), things (sensory events and perceptions), and concepts (ideas, notions, and thoughts). This categorization is based on the amount of abstraction involved in the subjects’ perceptions when using different methods. This approach provides a more clear view on what methods generate what type of data and with how much abstraction. For example, perceptual properties such as pitch and loudness require very little abstraction by the subject, whereas product-sound quality and content quality require a very large amount of abstraction by the subject [6]. This example illustrates further the notion that perceptions, which are objective and which are interpreted, are difficult to distinguish. However, depending on the method and what listeners use, one can collect data that are more/less objective. When changing the point of departure, one needs to know how these premises are different, in this case, between qualitative and quantitative points of departure.
2.3. Distinctions between a qualitative point of departure and a quantitative point of departure

In qualitative research, the validity of the research does not depend on the replicated outcome, rather the focus here is on the process in which the results were derived from: the way the results were obtained (the analysis) and the way the results represent the collected data. Two researchers will never produce the same results even if they are presented with exactly the same task because of their different philosophical stances and individual styles, creating different viewpoints of the examined phenomenon (research question) [10]. This is very different from quantitative research where the objectivity and the accuracy of the measurement results and replication of results are crucial. In qualitative and quantitative research, inter-subjectivity should be sought (see section 2.2), but there are differences between quantitative and qualitative research. This issue is highly debated and this discussion should be seen as an overview of the main differences.

The critical testing of inter-subjectivity on the results in social science are composed of two conditions according to Bergström [11 p. 113]:

1. In principle it is possible for anyone with suitable intellectual capacity and technical equipment to critically test the results; and

2. If several persons should critically test the results, all should be able to draw the same conclusions.

Bergström [11] makes the distinction that all social science results are able to fulfil condition 1 in practice although it is very difficult to achieve condition 2 in practice, but in principle it could be achieved. To fulfil condition 2, all the conditions in which the research was conducted need to be replicated; this can be done in theory, but it is very difficult in practice, especially in situation-based research, which is often seen in the social sciences. Therefore, it is very important for the social science results to fulfil condition 1, but it should also strive to fulfil condition 2 in principle [11]. It is important to know this distinction when comparing qualitative research with quantitative research. In more experimental-oriented research, as with perceived sound quality evaluation, both conditions 1 and 2 are more easily achieved because the researcher can control all parameters, an experimental approach not possible in the social sciences.

In this section, the major differences between qualitative and quantitative research were presented. In the following sections, a more detailed description of the research field in questions is presented.

2.4 Perceptual audio research

A significant part of the audio research methods has its foundations in the field of sensory evaluation research [12,13] and can be found, for example, in the natural sciences and psychophysics. All sensory evaluation methods have their own guidelines and techniques for response tasks, statistical methods, and interpretation of the results. The general approach of sensory evaluation is to evoke, measure, analyse, and interpret responses collected from subjects’ sight, smell, touch, taste, and hearing senses [12]. This general
attitude can be seen in Bech and Zacharov’s [9] criteria on scientific method used for perceptual audio evaluation. The following four criteria have been extracted from their work [9].

Objectivity\(^1\): Another experimenter should with the same circumstances, with the same stimuli, with a similar data set, and with the same statistical analysis come to the same conclusions during a new experiment and with new subjects.

The quantification of the impression: There is a direct relationship between an auditory sensation and attributes under investigation. Answers (numeric data) collected from scales can be tested statistically. The researcher assumes that there is a direct relationship between the testable statement and the answers the subjects provide (the dependent variable).

Communication: Using defined and specific attributes in the experiments clearly presents the communication between scientists, clients, and the public.

Scientific generalization: The conclusions can be generalized to a lager mass, based on a small representative population of subjects.

There are many research methods employed in perceptual audio evaluation research all with their own niche depending on the aim and purpose of the research. The methods can be categorized in at least two ways: the method’s level of abstraction [6] or the perceptual domain. The latter categorization will be described here and is the most frequently used. See Figure 1 for the different domains.

The physical domain deals with measurement of the physical audio signal and the sensory and affective domains deal with the auditory sensation of an audio signal. These three domains and their associated research methods will be presented in the following subsections.

\(^1\) This definition of objectivity is closely related to the reliability of experiment, one part of the trustworthiness of the experiment. The second part is the validity of the experiment.
2.4 Physical domain
The physical domain is highly objective in the sense that the researcher can accurately measure the physical nature of an audio signal such as a loudspeaker response, frequency content analysis, and reverberation time [4]. No subjects/listeners are needed in this type of measurement. There is also a branch of research that uses objective models/methods to predict or measure perception and audio quality. For example, Bradley and Soulodre provide an objective measurement of listener envelopment in multichannel surround systems [14] and the ITU-Recommendation BS.1387 provides a method to measure audio quality objectively [15].

2.4.2 Sensory domain
According to Lorho [4] the sensory domain can be seen as

\[\ldots\text{the elicitation of objective responses to the properties of a stimulus, as it is perceived by the human sense. ([4] p. 7)}\]

This domain can also be labelled perceptual measurement [9]. It is assumed that an auditory sensation is based on several auditory attributes and that the researcher can identify and elicit these attributes. The subjects’/listeners’ task is to quantify the impression of one or several of the elicited attributes during an investigation [9, 16]. For example, Gabrielsson and Lindstrom [17] and Toole [18] elicited attributes and used the attributes when investigating the perceived sound quality of loudspeakers.

There are many methods for eliciting the attributes and for using the attributes. The methods can be divided into two large categories: methods that separate sensation of an auditory impression and the verbalization and methods that assume that there is a direct relationship between the auditory impression and verbal descriptors used by the listener/subject [9]. The former method might draw on spatial impressions [19] or indicate direction of a sound [20]. Martens and Zacharov’s research [21] provide another example of a method in which the listener does not label a sensation, the Multidimensional Scaling (MDS) method. Here the focus of the investigation is on the individual differences/similarities of pairs of stimuli. The methods that fall in the latter category can investigate a specific auditory attribute of loudspeaker sound quality [18], descriptive analysis, or attribute elicitation using a repertory grid technique. Koivuneiemi
and Zacharov [22] used audio descriptive analysis and mapping (ADAM) and developed a common language (common attributes) to unravel spatial sound perception. Berg and Rumsey showed that a repertory grid technique is a fruitful method for eliciting individual attributes in spatial audio [1].

### 2.4.3 Affective domain

The affective domain methods measure the subjective and global perception of the auditory sensation [4] to quantify an “overall” and “general” impression of the auditory sensation. The listener is assumed to have an integrative state of mind when creating the overall impression. The listener should also form a single impression taking into account the individual attributes of the sound, the context, the mood, the expectation, the previous experience, and traditions [9].

Three tests are associated with the affective domain [9]:
- Preferences tests (paired preference tests and ranking based on preference);
- Acceptance tests (report the degree of liking); and
- Appropriateness tests (report the degree of liking with the context relationship) [23].

### 2.4.4 Response format

The methods employed in perceptual evaluation research can be categorized into three general sensory tests [12] based on their focus:

- Discrimination tests (the focus on overall differences among products);
- Affective or hedonic testing (measuring consumer likes and dislikes); and
- Descriptive analysis (specification of attributes).

Below are some examples of the methodologies and their accompanying response formats.

The discrimination tests can also use an ABX comparison [24] using a nominal scale or an AB comparison using an ordinal scale to decide whether two objects are equal. The AB comparison is used to decide whether a magnitude of a sensation of one object is higher/lower than another object [12,13]. Other examples of scales and methods used here are the ITU Recommendations BS.1116 [3] and MUSHRA [25], both of which use interval scales to show how much an object is greater or less than a non-equivalent object [9]. These methods [3,25] also show the magnitude of the differences. Affective or hedonic testing, where one tests consumer likes and dislikes using a partition scaling, divides the sensory continuum into equal intervals [9]. The descriptive analysis looks at how a product differs in specific sensory characteristics [12, 13]. Zacharov and Koivuniemi [22], for example, use an audio descriptive analysis and mapping experimental design to investigate the perception of spatial sound reproduction. The response format includes a 10-point scale using a one decimal place interval scale that includes positive and negative end words for each specified attribute. Twelve attributes were used and comprised of both spatial and timbral attributes.
2.4.5 Implications of perceptual audio research
Perceptual audio evaluation methods have several advantages since they aim at generalizing and predicting the results. However, there are several implications with the use of these methods. As they are often conducted in experimental conditions, this is not a major drawback; however, the methods are hard to apply in contexts other than experimental conditions where the researcher does not have control over the parameters. This limitation makes case studies very difficult.

In addition, it is even harder to arrive at a more elaborated understanding into how the subjects perceive different attributes, differences, and qualities, since the quantification of the impression does not allow the subjects to describe their impressions in detail. Several questions could be investigated to collect a more complete understanding of the subject’s experience:

What goes into an affective judgment?

What does the subject actually perceive more than the global quality, good/bad, liking or disliking?

What are the differences the subjects perceive in a BS.1116 or MUSHRA experiment?

It is important to note, as Berg and Rumsey [1] and Kjelden [26] discuss, that there are limitations in using established attribute scales. One limitation is often terms used are not consistently defined and have different meanings for different subjects. If a consensus has been established on the meaning of the terms used (e.g., by expert listeners), then the meaning of the term might not be valid for the rest of the world. Another limitation with using established attributes is the fact that it only gives an answer to the question posed [26]. To overcome these limitations, both Berg and Rumsey [1] and Kjelden [26] propose the use of individual attributes. Similarly, many researchers have used a mixed method approach for collecting both qualitative data as well as quantified data to understand the quantitative results even better.

2.5 Mixed methods research
Mixed methods research combines several methods to investigate a topic from several points of views. Strohmeier et al. [27] used a mixed method approach to evaluate multimodal quality perception of audio and video. The combinations of methods were preference mapping, sensory profiling, and a method of constructing a link between the preference mapping results and the sensory results.

Jymsiko-Pyykkö et al. [28] used semi-structured interviews (quality evaluation criteria) and a single stimulus method (quality evaluation) to evaluate audio and visual quality. The data analysis in the particular research used a combination of grounded theory analysis (finding patterns and building a theory from the data) [29] and quantitative statistical analysis (Bayesian modelling). And in [30] results form semi-structured interviews and grounded theory analysis were compared with quantitative data in order to investigate the relation between produced and experienced audio-visual quality.
Schulte-Fortkamp and Fiebig [31] also use a mixed method approach. They conducted a case study on evaluation of noise in a residential area using both interviews and acoustical measurements to investigate people’s attitudes, perceptions, and noise in an residential area. The analysis had a grounded theory approach. The results from the study showed that the grounded theory approach could explore, in detail, what influences the process of perception. It was also purposed that further investigation on the integration of qualitative and quantitative data should be conducted.

2.5.1 Implications of mixed methods research

These examples show that a mixed method research approach could be fruitful since one can collect different sets of data that can give information from different viewpoints on a particular topic. The latter work is interesting since this research is conducted “outside” the experimental boundaries, which indicates that a similar approach can be useful in a situation that is not fixed to a test laboratory. It is also interesting to see that two of the mixed methods shown have been used in sociology to collect more detailed information. When trying to understand audio perception, perhaps researchers should borrow from the humanities and social sciences.

2.6 Qualitative methods

Both the humanities and social sciences are broad fields of research, including sociology, history, and psychology, just to name a few disciplines. These fields employ a wide variety of qualitative research methods that are highly intertwined with their respective methodologies. The methodologies are important to understand when using qualitative research methods because each method has its own framework that the user must acknowledge.

Qualitative methods can be productive for perceptual audio research, because qualitative research focuses on an

"[...] attempt to understand the meaning or nature of experience of persons [...] getting out into the field and finding out what people are doing and thinking. ((32] p.11"

Qualitative methods are used to explore areas where little is known or much is known in order to gain novel understanding. It can also collect details about a phenomenon. The details can include feelings, thought processes, and emotions [32].

Anselm and Corbin [32] identify three major components to qualitative research (p. 11-12):

Step 1. Data collection.
Data can be collected from various sources: interviews, observations, documents, records, and films.

Step 2. Analysis of data.
Procedures used to organize and interpret the data. This includes conceptualizing, reducing the data, elaborating categories in terms of their properties and dimensions, and relating through a series of prepositional statements in which all are referred to as coding.
Step 3. Written and verbal reports. The results presented in the form of articles for books, scientific journals, or conferences.

To interpret/analyse the collected qualitative data, two main ontological standpoints and many methodologies of qualitative research come into play, phenomenology and hermeneutics, because of their predominant theoretical impact on issues concerned with understanding and interpretation of a phenomenon, human experience, or text under investigation. This methodology focus fits well with the departure point of this work presented in 2.2. Both of these vast methodologies will be presented in the following sections. Earlier research of perceptual evaluation has used grounded theory (see section 2.5), which is a method for deriving a theory from the data, and is not a methodology in the sense of phenomenology or hermeneutics. However grounded theory is approach that will also be presented because of its previous use.

2.6.1 Phenomenology
A phenomenology approach, which has it roots in a philosophical tradition, assumes there is a common essence or essences associated with an experience. The essences are the core meanings mutually understood through a phenomenon [33]. Moustakas [34] states that one can determine the underlying structures of an experience by interpreting the description of a situation where the experience occurs. Hence, by looking at the experience of several subjects, the phenomenon under investigation can become understandable to the researcher as well as to other uninitiated researchers.

Typically, phenomenology research collects data using interviews, “phenomenological interviews”. These interviews use three processes in design and analysis: bracketing, phenomenological reduction, and horizontalization. Bracketing (epoche) is a process where the researcher, before starting the data collection, becomes aware of his/her prejudices, viewpoints, and assumptions. This self-reflective approach helps the researcher avoid interfering with the way the structure of the phenomenon is seen. Phenomenological reduction is the process of reading and describing several times to comprehend and understand the text under analysis. The next step, horizontalization, is a process closely related to phenomenological reduction. In this step, the researcher organizes the data into themes or clusters to make the essences of the investigated phenomenon visible [33].

2.6.2 Hermeneutics
Hermeneutics looks at the human experience by studying the lifeworld\(^2\). Studying the lifeworld, the researcher can from a whole, including the description of experience [34]. Interpretation is a commonly used term in hermeneutics. The interpretation process or analysis process takes the form of a spiral called the hermeneutic circle/spiral. The researcher moves back and forth between individual elements (text segments) and the whole (entire text) [10]. This process is to make the intentions and meanings behind appearances in the text clear and understood [34] and to make one or several events understood in the context in which it is a part of [10].

\(^2\) Lifeworld: “all the immediate experiences, activities, and contacts that makes up the world of an individual or corporate life”. New Oxford American Dictionary, Retrieved April 13, 2012
An important point is that phenomenology and hermeneutics are closely related. In phenomenology, there is a subset called “hermeneutic phenomenology”. This subset sees the world or experience as a text that must be read [10]. This division between the different methodologies is not sharp and there can be some overlap between them.

2.6.3 Grounded theory
The main characteristic of grounded theory, as a method, is that the researcher should collect descriptive data before developing theories and hypotheses [10]. The focus is on generating a theory from the data and being grounded in data [29] using descriptions (e.g., interviews), conceptual ordering (e.g., coding), and theorizing (connecting concepts from the data). The researchers are encouraged to be creative in analysing the data. The term creative is determined by the researcher’s ability to name categories, ask stimulating questions, make comparisons, and extract an informative scheme that is grounded logically, systematically, and explanatory in a large amount of unorganized data [32].

To deal with large amounts of unorganized data, researchers use theoretical sampling. The researcher jointly collects, codes, and analyses the data. The coding process is called “open coding” and involves looking at the data and establishing segments consisting of behavioural actions and events, both observed and described. Constant comparison method is also used. This method consists of comparing one segment of data with another segment of data to look for similarities and differences. The similarities are sorted in a similar dimension, which are named. After some time, the dimensions become categories when no new dimensions can be identified [33].

2.6.4 Comparison of qualitative research
The two methodologies – phenomenology and hermeneutics – collect descriptive data. This follows the viewpoint of perceptual audio evaluations where the subject is the central measuring device\(^3\) in which data are collected from. There is a difference between grounded theory and phenomenology: grounded theory creates a theory from the data and phenomenology begins with a theory. This indicates that grounded theory could be more fruitful when no preconceptions are involved (in theory) and when exploratory research is needed where low amount of information is known.

With phenomenology, on the other hand, the researcher can investigate in detail the phenomenon by asking questions about the phenomenon and checking the results against a proposed theory in order to accept it or reject it. The phenomenology approach also resembles the departure point presented in 2.2 seen from an ontological view, where the objective world is obtained through our senses, seeing the objective world as a phenomenon that needs to be studied. Then it is feasible through a phenomenological approach to reach the objective world, at least in theory.

---

\(^3\) Central measuring device is a term use by ([9] p. 106) to describe the listeners’ function in a listening test.
Seen from an analysis perspective, the hermeneutics approach is unavoidable, since it can shed light on the intentions and meanings that inform the listener’s descriptions, which could unravel interesting results from a listening test. In addition, the interpretation process of the analysis involves reading/studying the data several times, in parts and as a whole, in order to understand it.

2.7 Summary of theory and background
The foundation of this work is based on the following: The only way to know anything about the objective world is to investigate how people perceive the world. Thus people make an interpretation of the experience of the objective world and this interpretation involves abstract thinking to a higher or lower degree.

Perceptual audio evaluation methods can predict and generalize results. The research is experimental-oriented, making the application of the methods difficult in real-life/non-experimental conditions. Moreover, since the experiments conducted are often stringently controlled to avoid bias, no other information about the perception can be collected than the answer to the question under investigation.

A qualitative approach is the opposite. It is situation-based research that collects a large amount of detailed information about a research topic. This can be done either by posing questions/theories before data collection starts or by deriving them from the collected data. However, there are several methodological implications when using a qualitative method in a field where quantitative praxis is common, such as perceptual audio evaluation, and the researcher should be aware of these implications.

A mixed-method research employs different points of view incorporating both qualitative and quantitative methods to investigate successfully a research topic, both outside and inside the experimental conditions. In addition, it can collect detailed information and generalize results and methods.

3. Research question and focus
3.1 Motivation
Perceptual audio evaluation methods have their merits in experimental conditions, where the researcher has control over variables and parameters under investigation. These methods, however, are not applicable in non-experimental and real-life situations where the researcher cannot control all variables, control what is happening, or guide participants to facilitate data collection, all primary concerns for evaluating perceptual audio methods. To evaluate perceptual audio quality in non-experimental and real-life situations, another method has to be applied. This raises the question: Can a qualitative method be applied successfully when evaluating perceived audio quality?

A phenomenological approach is very interesting since its ontological and epistemological foundation resembles the goal outlined by perceptual audio evaluation: determining the underlying structure of experience. There is also evidence that mixed method research could be applicable because it can be used in experimental and non-
experimental conditions. Furthermore, data collection can be applied before, during, and after an experiment or non-experimental situation.

A qualitative method also gives another aspect to a perception in comparison to a quantitative method. Using a mixed method approach, a researcher can collect two aspects of information about a research topic. This allows for an understanding of the perception that is broader than a single quantitative or qualitative research method could provide.

3.2 Aim
This study investigates whether a phenomenological research method can be applied with success in the field of perceptual audio evaluation. The productiveness of a qualitative method is shown in the form of what type/types of information can be elicited and the usefulness and the relevance of the results elicited by the method. By looking at this information, an understanding of what the method can be useful for can be reached.

3.3 Objectives
To accomplish the aim of this study, the main question (i.e., can a phenomenological method be successfully applied in perceptual audio evaluation research?) was broken down into sub-questions:

- What information can this method generate?
- Is the information/results collected relevant\(^4\) for the conducted study?
- Can the information/results elicited be useful\(^5\)?

To answer each sub-question, the qualitative method was incorporated into three different studies (Papers 1, 2, and 3). Each study used a different design. This allows for the qualitative method to be examined in various conditions. It also gives insight into the relevance and usefulness of the results, and what type of information can be elicited for that particular study. In Figure 2, the study design is illustrated.

Each study used a different design and had a different purpose and aim. The designs used were a non-experimental study, a study conducted in experimental conditions using a mixed method, and a study conducted in experimental conditions. Each study’s research topic is explained below.

**Paper 1 (Non-experimental study):**
Evaluate the sound and video quality of Master classes in classical music performed over the public Internet using a qualitative method.

---

\(^4\) The term *relevant* is defined as [adjective], closely connected or appropriate to the matter at hand. *New Oxford American Dictionary* (2012-05-22).

\(^5\) The term *useful* is defined as [adjective] able to be used for a practical purpose of in several ways. *New Oxford American Dictionary* (2012-05-22).
Paper 2 (Qualitative Method applied in addition to a standardized listening test): Investigate what additional data can be collected about the subjects’ perception using a qualitative method in a BS.1116 and MUSHRA designed listening test.

Paper 3 (Study conducted under experimental conditions): Investigate what information can be collected using a qualitative method as the only data collecting method in an experimental condition (preference ranking listening test).

Evaluation of a qualitative method applied in perceptual audio evaluation research

<table>
<thead>
<tr>
<th>Study 1: Qualitative method applied in a non-experimental situation (Paper 1)</th>
<th>Study 2: Experimental condition using a mixed method approach: The qualitative method applied to a standardized listening test (Paper 2)</th>
<th>Study 3: Qualitative method used in an experimental situation (Paper 3)</th>
</tr>
</thead>
</table>

Figure 2. Study design

4. Summary of papers

4.1 Paper 1
In Paper 1, the main goal was to evaluate perceived sound and video quality in a real-life/non-experimental situation – master classes of classical music taught via the Internet. To accommodate for this real-life situation in which the study was conducted, questionnaires and interviews were used as data collection methods. The questionnaires were analysed by looking at trends/major themes in the data and by using meaning categorization (inspired by Verbal Protocol Analysis). The analysis of interviews used meaning condensation to elicit the subjects’ perception. Both analysis methods are based in phenomenology and hermeneutics. The results from the questionnaire gave information about the following characteristics:

- Sound quality: involving statements about the clarity and naturalness of the sound and the easiness of understanding the teacher.
- Asynchronicity between video and audio: involving statements about delay between sites and delay between picture and video.
- Teaching-related information: e.g., the teaching was excellent.
- Communication-related information: e.g., the difficulties in student teacher communication.
The meaning categorization gave a numerical overview of the general attitudes of the participants on the perception of sound, video, and teaching quality. Positive comments about teaching and sound quality were given the highest numeric value. Teaching, sound quality, and sound- and video-related were the most common comments.

The interviews revealed several problems or issues. Perceived audio and video quality contained, for example, information about the subjects’ descriptions of their perception of sound and video quality by comparing it to known media, e.g., life-like MP3 and video quality better than a YouTube clip. The equipment’s limitations and student’s limitations could be separated from a teacher’s point of view. Perceived problems and potential problems with the equipment and teaching were also noted, for example, the delay between audio and video makes communication with students difficult, small details lost, and difficulty distinguishing the students’ playing technique. However, the equipment did allow for controlling tempo, intonation, articulation, and phrasing. Perceived differences and similarities with teaching master classes over the Internet were also noted. There were differences when correcting playing technique and creating a teacher/student relationship. However, using the equipment to meet with students seemed to be the same as meeting with students in, e.g., a teacher’s office.

The results collected were detailed about perception of audio and video quality. The results also showed that the interaction between the subjects (teacher/student) was also a factor involved in creating the total judgment of quality.

4.2 Paper 2
Paper 2 had four main goals: 1) to investigate the amount of transparency reached for higher bit rates in DAB+; 2) to compare Swedish Radio FM and DAB+ chains in order to see how they relate to one another in quality; 3) to map the audio quality of the DAB+ and FM chains to the audio quality scale in the MUSHRA method; and 4) to investigate what further information can be elicited by applying interviews after the listening test. The fourth goal is the emphasis here. Two listening tests were conducted, one BS.1116 method and one MUSHRA method. After completion of both tests by a subject, an interview was conducted with each participant. A total of 25 interviews were conducted. Four of these interviews were randomly selected for further analysis using meaning condensation with a phenomenological point of departure. The main questions here were on what and how the subjects perceived the degradations. The results gave the following information:

• What degradations were perceived, relating to, e.g., frequency response changes and phase problems.
• Which of the degradations were prominent, e.g., perception of stereo image and frequency.
• Which of the degradations were hard/easy to perceive – easy perceived degradations such as noise levels and panned transients and hard degradations to perceive such as complex stimuli containing no reference points.
• What degradation gave low and high scores – low scores, e.g., large differences between the stimuli and the reference – and how much the degradations were audible in combination with how much they affected the results.

In addition to these results, information about the subjects’ comparison and evaluation strategies and perception of the scales used was collected.

4.3 Paper 3

Paper 3 addresses the issue of what goes into an affective judgment. To investigate this issue, a listening test experiment (researcher controlling all parameters) was designed and carried out. The objective of this research was to investigate the method and results to examine what information can be elicited from it. The subject’s task was to rank stimuli according to preference. The data collecting methods were interviews (two interview parts per subject). One initial interview was conducted before the test to collect information about each subject’s listening habits, attitudes, expectations, mood towards the test, and preference of music genre. A second interview was conducted during the listening test to collect information about what the subjects listen to, what differences they perceive, how or if they associate the perceived differences to something, and why they ranked the stimuli in this particular way. The analysis method used was meaning condensation. A total of three interviews were completed during the listening test. The results show information on several aspects of what goes into an affective judgment.

• Method. What way does each subject listen after differences? The results showed different strategies in listening after differences, e.g., listening to the whole, first impression, specific parts, listening beyond the music, and how well the verbal information came cross.

• Perceived differences. What differences does the subject perceive? The differences could be divided into two categories, direct differences and associative differences. Direct differences involved, e.g., change in frequency content between stimuli, sharpness of consonants, changes in levels between instruments, changes in breathing sounds, and how well the verbal information came across. The differences based on association involved the subject to associate certain differences to pervious experience, e.g., radio commercials, how news or pop music “usually” sounds, or being placed at different locations in the mix.

• Ranking. How does each subject make rankings according to their affective judgment? The ranking process was also divided into two main categories, ranking based on a combination of associated differences and direct differences, from above, and ranking based on association, e.g., ranking based on where the subject would listen to the stimuli (such as at home), one specific ranking, and doing a radio show, another ranking. From this part, an indication is given that listeners relate their associations to their background, e.g., the subject ranking in two separate ways as above also works with radio drama/radio education experience.
An interesting finding emerged when looking at the initial interview and the ranking order from one of the subjects. One subject ranked the stimulus with the lowest bitrate as most preferred, while the same subject’s listening habit was predominantly listening to audio with low bitrates. Thus a causal relationship between preference and listening habits was evident although further research on this topic is needed.

5. Summary of Results

5.1 Types of elicited information
In summary, the information elicited from the qualitative method is related to the research topic of each study. The types of information elicited by the qualitative method from each study are summarized in Table 3. It can be seen from the results that the information elicited is not limited to the research question only and include other aspects:

• Teaching- and communication-related issues [P1];
• Differences and similarities between distance music teaching and regular teaching [P1];
• Interaction (communication) between different parts is a large part of the total perceived quality [P1];
• Perception on the usability of the scales, BS.1116-1 and MUSHRA, in the study [P2];
• Subjects’ attitudes towards the investigated study [P1 and P3]; and
• Strategies/methods used by the subjects in the study [P2, P3].
Table 3: Summary of elicited information from studies.

**Types of elicited information**

<table>
<thead>
<tr>
<th>Publication 1</th>
<th>Publication 2</th>
<th>Publication 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Info. about audio and video quality perceived by the subjects (users).</td>
<td>- Degradations perceived.</td>
<td>- Initial expectations/mood etc. of the subjects before the test.</td>
</tr>
<tr>
<td>- Teaching related info.</td>
<td>- Degradations that were prominent.</td>
<td>- How the subjects listen for differences.</td>
</tr>
<tr>
<td>- Communication related info.</td>
<td>- Degradations that were hard and easily perceived.</td>
<td>- What differences were perceived.</td>
</tr>
<tr>
<td>- Attitudes towards distance music education. Perceived problems and possibilities with the equipment and teaching.</td>
<td>- How degradations influenced high and low scores.</td>
<td>- What the subjects associated with the differences.</td>
</tr>
<tr>
<td>- Perceived differences and similarities in teaching.</td>
<td>- Perception of the scales used.</td>
<td>- How the subjects ranked and incorporated the perceived and associated differences in the ranking process.</td>
</tr>
<tr>
<td>- Interaction is a large part of the perceived total quality.</td>
<td>- Evaluation strategies used by the subjects.</td>
<td>- How the subjects incorporated previous experience and background info. in the affective judgment, both when perceiving differences and ranking stimuli.</td>
</tr>
</tbody>
</table>

5.2 Relevant information for the study

Paper 1 answers the research question that involves the evaluation of sound quality and video quality. In addition, paper 1 reveals other aspects about perceived quality. In paper 2, several types of information were relevant for the study – e.g., perceived differences, the degree of differences, and whether the differences are easy or hard to perceive. Paper 3 shows what an affective judgment is composed of and how the subjects carry out their affective judgments. This evaluation involves listener experience/background, direct differences, and associations with differences based on association and ranking based on differences.

5.3 Usefulness of the information

The material below outlines the usefulness of the studies.

- The information gathered increases the knowledge of what is being investigated as well as adding new information previously unknown [P1, P2, and P3].
- The information gathered increases the accessibility to the subjects’ methods when conducting a task, e.g., a listening test or affective judgment [P2 and P3].
- The information gathered about the method shows that the method can be useful in a non-experimental context [P1], in an experimental context [P3], and in a mixed-method experimental context [P2].
6. Discussion

6.1 Relation to quality aspects
The studies investigate different qualities: Paper 1 investigates the standard definition of quality (definition 3) used by standards [2] and P1, P2, and P3 investigates the “an event” quality definition (definition 4). The results of these investigations indicate that this particular qualitative method can bridge the gap between the regular definition of quality (definition 3) and the more subjective definition of quality (definition 4).

It is also interesting to note that the studies incorporate several of the different abstractions levels presented by [6] (Table 2). The elicited information from the studies involves, e.g., perceptual properties (Auditive Quality), speech intelligibility (Aural-scene Quality), and assigning of meaning and dialog quality (Aural-communication Quality), further enforcing the notion that the information elicited is rich in details.

6.2 Framework of Validity and reliability
Qualitative research should hold up to critical testing in order to make it valid and the researcher should strive to be as objective as possible (see section 2.3). In addition, it is very hard to come to the same results in different studies in qualitative research because the studies are situational. Each study presented here has fulfilled condition 1, posed by [11]:

*In principle it is possible for anyone who has suitable intellectual capacity and technical equipment to critically test the results.* ([11] p.113).

As such, qualitative studies are needed; however, such studies are difficult with respect to condition 2:

*If several people were to critically test the results, all should be able to draw the same conclusions.* ([11] p.113).

As stated before, these are major differences between a qualitative and quantitative approach. In quantitative research, one can easily reach condition 2, e.g., one can create the same conditions in which the experiment was conducted in, elicit a random sample out of the same target population of individuals, and use the same statistical analysis to interpret the same sets or new sets of data. The requirements of quantitative research – objectivity in the sense of reliability and repeatability, quantification of the impression, communication, and generalization [9] – cannot be applied to qualitative results. However, this does not mean that anything goes when it comes to qualitative research.

Qualitative research focuses on the individual experience and how the individual perceives this experience. The subjects are picked because of certain criteria, e.g., experience of the phenomenon under investigation, which is called purposeful sampling. In a qualitative approach, the different viewpoints/aspects collected from the subjects give an understanding of the phenomenon under investigation. The qualitative results should be critically tested on their own terms and the research should address these main concerns: should fulfil condition 1 and strive for condition 2, should focus on the
analysis, should represent the data collected, and should provide a different view or aspect about a phenomenon from study to study [10] and between each subject.

This research has focused on phenomenological interviews and analyses. The method has been judged and critically tested based on qualitative scientific requirements. The studies have been the means for showing the usefulness of this method. The method has been critically tested in various forms, from non-experimental and experimental contexts. The methods used in each study have been described in a scientific manner that fulfils condition 1, and the analysis methods and the results have been scrutinized in each paper to reveal the applicability of the methods.

6.3 Interviews and analyses
One underlying assumption of this work has been that there is a direct relationship between what is perceived and that is verbally said. With this assumption a difficulty arises: terms used could have different meanings for different subjects [1]. To adjust for this difficulty, the accuracy of interpretation of the interview data has to be high. This inflicts further difficulties such as interpretation accuracy. The phenomenological approach uses three steps to elicit results with accuracy: bracketing, phenomenological reduction, and horizontalization. As described previously, the bracketing is the process of having an objective approach towards the data collection and analysis. Thus the researcher avoids imposing any leading questions or any personal viewpoints when collecting and analysing the data. This is done using non-leading/open-ended questions in the interviews that allow the subject to respond to an interpretation made by the interviewer of a statement during the course of the interview. In the analysis, the bracketing becomes a bit harder. The interpretation process of the interview data, post interview, involves phenomenological reduction and horizontalization. These steps involve abstraction of the data, on several levels, to understand it. In direct results where the interviewee states a particular difference (e.g., too much low frequencies), a low level of abstraction is needed, but in more elaborated results such as ranking procedure or association to differences, the hermeneutical approach comes into play. In this case, the researcher reads parts of the transcribed interview and the whole several times and uses his/her knowledge to understand what the subject meant by a particular statement and this involves a higher degree of abstraction. This approach facilitates the collection of accurate data about a perceived phenomenon.

A possible critique to this method, seen from a quantitative point of view, is the need for consensus between researchers. From a qualitative point of view, this criterion is not as important because if the results and analyses are thoroughly described, one should, in principle, be able to, with the same reasoning and the same data, draw the same conclusions.

6.4 Information elicited from the studies
The qualitative data elicited in earlier research [28, 30, 31], which have been used in qualitative methods, and the data presented here [P1-P3] have similarities. The similarities indicate how a perceptual evaluation is constructed. [31] concluded that one could see in the results that sound evaluations depend on the social and cultural structures
in which an individual is imbedded. Knowledge about explored interdependencies could be used in future research to increase awareness of existence of environmental stimuli that can be harmful. [30] elicited several audio-visual quality experience factors from the qualitative data and showed that there is no 1:1 ratio between produced quality and experienced quality. [28] elicited even more detailed information about experienced quality factors from the qualitative data. The factors could be divided into high-level factors (content factors and usage factors) and low-level factors (audio, video, and audio-visual quality) and the connections between the factors. Paper 1 gives information about audio and video quality as well as teaching and communication factors that are involved in the quality judgment. Paper 2 shows what and how the differences are perceived and paper 3 gives detailed information about the composition and construction of an affective judgment.

The perceptual research methods BS.1116 or MUSHRA [3,25] are limited to their scales. By using interviews in addition to these tests, as in paper 2, an understanding of how the scales are used and what subjects actually perceive can increase the knowledge of what is going on during the test. It also gives the researcher the ability to compensate when errors occur. One could argue that a conversation with the subjects after each session would suffice, but by using a qualitative approach the data can be scientifically analysed.

The data/information collected from an interview or questionnaire depends on the type of questions stated. If the researcher does not follow up on interesting comments, the data will be limited. Showing a relationship between questions and answers is needed, so experience and training in interviewing can be essential for the researcher collecting the data.

These qualitative methods can generate detailed information about an experience and a perception relevant to the study design. The information can be used to better understand the user experience of perception. This information can be highly valuable for any institution, company, or entity trying to understand and implement a new technology. In addition, the information generated from these types of qualitative methods can give insight from a relatively small population of subjects (e.g. from one to five) that can then be used to create larger quantitative experiments on different aspects of that particular experience.

6.5 Using the qualitative methods
In conclusion, this study has shown that a qualitative approach using phenomenological interviews and analyses can elicit results that are valuable for the field of perceptual audio evaluation during experimental and non-experimental conditions. Thus this approach gives detailed information about the subjects’ experiences and perceptions both as a single methodology and in combination with quantitative research methods. This methodology allows the researcher to evaluate the subjects’ perceptions and extract what the perception is composed of and how it is used. This study also identifies several implications with using a qualitative method when studying perceptual audio:
• Knowing the ontological stance and departure point of the qualitative method to analyse and elicit information from the data; and
Knowing the criteria a qualitative method is judged upon and how detailed the method and results are described. In principal, it is possible to achieve the same result in the same setting.

### 6.6 Original contribution

This thesis has applied qualitative research methods in the field of perceptual audio evaluation research. The method has been used in several different conditions to investigate its usefulness and relevance to the field. The investigation shows in detail several implications (ontological, epistemological, and methodological) associated with using this method in the perceptual audio evaluation field in which the researcher has to be aware of when using the qualitative method.

### 6.7 Further work

Seen from a quantitative point of view, this method needs further evaluation:

- Enhancing the analysis process by adding a second analysis conducted by a second researcher before comparing the results from each analysis in order to check for consensus of the results and to further enhance the inter-subjectivity of the interpretation of the results; and
- Using participants with a larger age span or background so as to collect answers from a large population of individuals, an approach that could give information about a broader view of perception.

Seen from a qualitative point of view, further research can be done in the form of conducting each study once more with new subjects in order to see the phenomenon in a new perspective so as to add information about perception in these studies.

The method can also be applied in other areas of research such as interaction-based applications and in an evaluation situation where the subjects cannot be interrupted such as when evaluating the interaction quality of an audio-visual system where subjects are interacting with each other via a piece of technology or when perceived quality of different graphical user interfaces (GUI) is evaluated where subjects are occupied with a particular task involving the GUI. The method can also be applied to evaluating the user quality, seen from the subject, of a listening test. This can facilitate a change in the design that could improve the listening test.

### References


3. BS.1116-1; Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems; *ITU-R Recommendations* (1994-1997)


23. Beresford, K; Ford, N; Rumsey, F; Zielinski, S. Contextual effects on sound quality judgments: listening room and automotive environments. *Presented at the 120th AES convention*, 20-23 May, Paris, France, 2006


30. Jumisko-Pyykkö, S; Reiter, U; Weigel, Chr, Produced Quality is Not Perceived Quality - A Qualitative Approach to Overall Audiovisual Quality; *3DTV Conference*, May 7-9, 2007


Paper 1:
A qualitative approach to evaluation of perceived qualities of audio and video in a distance education context

A qualitative approach to evaluation of perceived qualities of audio and video in a distance education context

Dan Nyberg\(^1\) and Jan Berg\(^1\)

\(^1\) Luleå University of Technology

Abstract This study presents a qualitative method for collecting and analysing data to describe audio and video quality. Used in the social sciences, arts, and humanities, this approach relies on phenomenology and hermeneutics and uses interviews and questionnaires to assess the audio and video quality of master classes in classical music taught via the Internet. Although this study is only exploratory, it provides evidence that the method could successfully be used to gather descriptions of perceptual qualities.
1. Introduction
There are numerous contexts in which evaluation of audio quality may be performed and several evaluation methods have been developed over the years. Although each method has its own merits, specific methods may be more suitable in specific contexts. When a new context emerges, these methods need to be evaluated, refined, and considered. Existing methods, however, mainly rely on experimental data that comes from carefully manipulated situations to control experimental variables. Exclusively relying on experimental data may not actually reflect the reality that the experiment aims to represent. Relying only on experimental data can be an issue when both audio and video qualities are of interest and/or when other quality aspects are involved. In addition, most of these methods work by quantifying the listener’s experiences, an approach that limits the type of data collected.

This limitation is imminent in the context of distance education over the Internet [1], including music education (MED), where research has increased over the last ten years [2,3]. Typically, a MED teacher located at one school teaches a student located at another school via the Internet, often by means of a video conferencing system [2,3]. Research efforts have been made to establish a transparent system for connecting two physical locations via a shared electronic virtual space for musical practice [4]. However, few studies specifically focus on how MED students and teachers perceive the audio quality of a particular system. Woszczyk et al. [4], for example, conclude that more aspects than audio quality are needed when evaluating such a system.

This paper presents a selected qualitative method for data collection and data analysis that is commonly found in the social sciences, arts, and humanities and applies these methods to evaluate audio and video quality as it is perceived by teachers and students in a non-experimental distance education setting (specifically in a real-life classical music course conducted via the Internet). The terms audio and sound quality will be used interchangeable during the course of this paper. The term non-experimental means that the researcher has no control over the parameters such as choice of participants, where the study took place, transmission quality between the locations, and choice of equipment. The term real-life means that the lessons in the classical music course are authentic lessons in which the researcher does not intervene. These qualitative methods are used because the study was conducted in non-experimental and real-life conditions. Although the results from such a study may be used to evaluate the quality of audio equipment, this issue was not the target of this study. Rather, this study intends to investigate what type of data a researcher can collect by using these qualitative methods and to show the applicability of these methods in the field of perceptual quality.

2. Earlier research on perceived quality
As stated in the introduction, MED involves more than one perceptual quality (e.g., audio and video). Previous research on the quality of perceived audio, which is one of the modalities under investigation, has been successfully evaluated. These data collection methods often deal with rating scales [5,6,7,8,9] and descriptive analyses and use the subject’s own vocabulary to evaluate the perceived audio quality [10]. There are also a variety of methods designed to evaluate perceived video quality. These data collection methods range from scales [11,12] to interpretation-based estimations of image quality [13] or both [14]. These approaches, however, only look at one modality.

In multimodal research, the focus shifts since both audio and video are involved in a MED situation. In the multimodal field, several approaches are used to evaluate multimodal quality.
Some of the approaches use a mixed-methods approach where both numeric and descriptive data are analysed. For example, Strohmeier et al. [15] used a mixed-methods approach to evaluate multimodal quality perception of audio and video. They used a new method, Open Profiling of Quality, to understand quality perception. The Open Profiling of Quality contained three parts: a method they call psychoperceptual evaluation (the subject uses rating scales to evaluate perceived audio quality); sensory profiling (collecting individual quality attributes); and external preference mapping (constructing links between the psychoperceptual result and the sensory profiling). Using analysis of variance (ANOVA) and Principal Component Analysis (PCA), they analysed psychoperceptual evaluation and sensory profiling, respectively.

Jumisko-Pyykkö et al. [16] used a qualitative approach to evaluate audio-visual quality and to test quality evaluation (single-stimulus method) and quality evaluation criteria (semi-structured interview). The data analysis combined a qualitative analysis based on grounded theory and a quantitative statistical analysis, Bayesian modelling. Jumisko-Pyykkö et al. [17] applied an entirely qualitative approach (semi-structured interviews) and compared the qualitative results with quantitative results in order to understand the relationship between produced and experienced quality in the context of interactive audio-visual application systems. The analysis used here also followed the grounded theory approach.

Other researchers have employed single-method approaches [18,19,20,21] in multimodal quality research. Beerends and De Caluwe [18] tested the influence of video quality on perceived audio quality (and vice versa) using a nine-point absolute category rating scale. The data analyses were made using ANOVA. Bech et al. [20] analysed the interaction between audio and visual factors in the context of a home theatre system. The experiment collected numeric data from rating scales using defined attributes with anchors and the data were analysed using ANOVA.

All methods presented above have shown to be well suited for the intended purposes and delivered valuable results; however, these methods are designed for experimental or laboratory contexts and some of the methods could be hard to adapt to a real-life situation (non-experimental situation) such as the MED situation. One cannot control all parameters involved in an MED situation, contrary to the experimental research presented above. Clearly, laboratory testing cannot completely account for the real life situations found in distant learning MED, a situation that is relatively unstudied and may require new methods to develop a complete picture of its usefulness.

3. A qualitative approach

When considering a suitable method in a relatively new, non-experimental, and multi-modal context, such as master classes in classical music over the internet, another foundation may be required that uses qualitative methods that focus more on describing and understanding the experience than on quantifying it. This approach makes it possible to adapt and employ methods used in other fields, in particular methods used in the humanities, social sciences, and psychology. Viewing a research area from a different research foundation can been fruitful; for example, Blauert and Jekosch [22] took a perceptionism view of the world, a view that relies on the belief that all knowledge is based on sense perception. This view allowed them to create a layer model of audio quality.
The following sections present the implications of using a qualitative approach and points of departure for qualitative research as well as descriptions and applications of the data and analysis methods.

3.1 Implications of using a qualitative approach
When researchers use qualitative methods, they need to consider the possible limitations of such an approach. In qualitative research, the question of validity focuses on the employment of the “reduction” process (analysis) that leads to a result and not whether the data can be replicated, a common view issue in the field of audio quality research. Thus qualitative research focuses on how clearly the analysis is conducted and described by the researcher. The important point of qualitative research is that two researchers do not produce exactly the same results when faced with the same task. This poses no problem, however. That is, competent and skilful researchers produce results easily recognized by other researchers. Each exploration/study should bring a different perspective on the phenomenon under study and each perspective creates a more comprehensive picture of the phenomenon [23].

3.2 Qualitative points of departure
Empirical phenomenology and hermeneutics, approaches frequently used in psychology [23, 24], are suitable points of departure for evaluating audio and video quality in a distance education context such as MED conducted via the Internet. Empirical phenomenology has the following goal:

To understand the general psychological meaning of some particular human way of being-in-a-situation [...] through a number of descriptions of this way of being-in-a-situation from people who have lived through and experience themselves as so involved. [23] (p. 40)

Moustakas, agreeing with Tesch [23], states that one can determine the underlying structures of an experience by interpreting the description of a situation where the experience occurs [24]. As for hermeneutics, Moustakas summarizes it this way:

[R]eading a text1 so that the intention and meaning behind appearances are fully understood. [24] (p. 9)

This view agrees with Tesch’s explanation of hermeneutics [23]. Hermeneutics looks at human experience by studying the lifeworld2 in order to form a whole, including a description of experience. Interpretation of an experience shows what is hidden behind an phenomenon [24]. This understanding means that an empirical phenomenological and hermeneutic departure may help evaluate audio quality, video quality, and their interaction in order to extract, interpret, and describe the experience of the persons involved in a MED situation.

Other researchers have already applied grounded theory to evaluate the perception of audio-visual quality [16, 17] as a way to understand the characteristics of multimodal quality. Grounded theory, an approach that is often found in sociology, generates a theory from data [25,26]. When conducting grounded theory research, a researcher should collect data before developing theories and hypotheses rather than developing theories and hypotheses before collecting data [23].

\[1\] The term “text” in a hermeneutic sense is very wide and incorporates human action, interview transcripts, and texts [23].

\[2\] Lifeworld is a translation from German, Lebenswelt: "All the immediate experiences, activities, and contacts that make up the world of an individual or corporate life" (New Oxford American Dictionary).
Grounded theory has similarities to the proposed approach above, as we want the data to give information to us about a phenomenon (in this case audio and video quality and interaction between teacher and students), but our purpose is not to generate a theory or a hypothesis about the phenomenon; rather our focus is to describe the perception of what is experienced in this specific context.

3.3 Collection of data

The two data collection methods used, questionnaires and interviews, can be applied post the MED situation, thus they do not influence student-teacher interaction. Both these methods are often found in qualitative research. Interviews are used in phenomenology research and to some extent in hermeneutics [23, 24, 27]. Compared to numeric-based data collection methods, both questionnaires and interviews can provide detailed descriptions of how audio and video quality are perceived, separately and together. Similarly, questionnaires have been used in earlier studies [28,29,30,31] with good results. Open-ended questionnaires can gather many responses from the subjects and give an overview of the field under study. Of course, questionnaires can contain questions where the responses are quantitative, e.g., by using scales, lines, checkboxes, etc., but such approaches are not considered here as they represent a methodology outside the scope of this paper. Interviews provide insight into how subjects experience the learning environment – audio, video, and student-teacher interaction. This approach also allows more flexibility in data collection because follow-up questions can be asked [27], which is not possible with questionnaires. Collecting descriptive data from interviews could be beneficial when the research field is relatively unexplored [16]. The following sections show how these data collection methods are applied in a MED context.

3.3.1 Using questionnaires

Two sets of questionnaires were distributed during several MED sessions. Questionnaire 1 is a brief questionnaire containing a set of open-ended questions. Based on the answers collected from the first questionnaire, not presented here, a second more elaborated questionnaire was designed (Questionnaire 2), see appendix 1, with more specific open-ended questions. The new set of open-ended questions was designed to encourage the participants to reflect on whether they perceived the audio quality and video quality as good or bad and why they perceived the qualities in these particular ways. The questions were designed to capture a comprehensible view of the participant’s perception of the overall quality of the MED system and the participant’s experience. Data from Questionnaire 2 were systematically analysed (described below).

3.3.2 Using interviews

As a next step, personal semi-structured interviews with the participants in the MED situation (students and teachers) were used to further shed light on the results from the questionnaire and on perceived quality. An interview guide with open-ended questions was used to obtain vital descriptions of the users’ experience [24]; see Appendix 2 for the questions used in the interview guide.

3.4 Analysis of the collected data

A phenomenological and hermeneutical approach is used as the point of departure for the analysis. The analysis focuses on describing the user’s perception of audio and video quality as well as the interaction between teachers and students. This section presents the data analysis methods.
3.4.1 Using questionnaire analysis

An ad-hoc analysis is used on the data and includes theme analysis and meaning categorization analysis [27,32]. In the theme analysis, themes are searched for in the collected data [32]. The theme analysis is accomplished by first listing all the collected answers under each question and reading all the answers for similarities and common answers [19]. This part of the analysis is analysed hermeneutically.

The meaning categorization analysis is an analysis that is inspired by Verbal Protocol Analysis [33,34]. Each answer is coded and counted according to its identified properties. The verbal protocol analysis requires an algorithm or description on how to handle each verbal unit [33, 34]. Berg [33] categorised the data into descriptive features and attitudinal features. Each category was then divided into two subsets. For descriptive features, the units were divided into unimodal (only audio modality) and polymodal (other sensory modalities). The attitudinal features were divided in the same way but into emotional/evaluative attitudes or attitudes related to naturalness. The current analysis, however, differs from Berg’s approach: each answer to the open-ended questions for both its descriptive and its attitudinal features was analysed as each answer may contain both. Both the descriptive and attitudinal features of each answer were interpreted based on the context (each question) from which they were taken. The sorting processes were conducted in the following way: Sorting 1 – each answer was sorted after its interpreted descriptive feature using the labels below; e.g., if the answer were related to audio quality, it was given a sound quality-related label (Sqr); and Sorting 2 – each answer was sorted after its interpreted attitudes, i.e., positive, negative, both, or blank. The following labels were used when sorting each answer into its respective topic and attitude:

**Sorting 1.** Descriptive features: Sound quality related (Sqr), video quality related (Vqr), sound and video quality related (Sqr&Vqr), communication related (Com) (“communication” here refers to speaking and interaction among the participants), teaching related (Tch), technology related (Tec), teaching and communication related (Tch&Com), and diverse statements (Div).

**Sorting 2.** Attitudinal features: positive (+), negative (-), positive and negative (+ & -), and attitudes that contained no positive or negative statements (blank).

Consequently, sorting 1 in the meaning categorization analysis provides an overview of the number of answers on each perceived quality. The second sorting (Sorting 2) provides an overview of the answers’ attitudes (sorting 2): positive, negative, both, or blank (general statements).

3.4.2 Using interview analysis

The interview data were analysed using an approach commonly used in phenomenology studies [23, 27], meaning condensation [27]. One or several persons can do this type of analysis. When looking at the data from a phenomenological point of view, the interviewer and/or researcher must set aside his/her pre-understanding of the phenomenon to obtain objectively rich and clear descriptions of the phenomenon under study [27]. This is referred to as “bracketing”. The bracketing started during the design of the interview questions by formulating open-ended questions to facilitate the subjects’ own descriptions of their experience and continued throughout the analysis process.
Meaning condensation concentrates the uttered meaning into more essential meanings and contains five general steps: Step 1 – reading the interview to obtain an overview of its content, establishing a sense of the whole; Step 2 – creating units of meaning (the answers) as the interview subject expresses them; Step 3 – creating themes that dominate the units of meaning; Step 4 – asking questions to the units of meaning based on the research purpose; and Step 5 – creating a summary of the interviews’ central themes and presenting them in one descriptive statement per interview [23, 27]. The steps of the interview analysis used the original method as a baseline for the analysis. The steps used in our analysis are listed below:

1. Transcribing the interview (transcription methods used were structured in colloquial language [27]);
2. Creating units of meaning of each transcribed answer;
3. Creating themes;
4. Sorting the units’ answers into corresponding themes; and
5. Summarizing each theme in the text for each interview.

All steps of the evaluation method are presented in a block diagram in Figure 1. Each step (block) is followed by its result.

4. Evaluation of perceived qualities of a master class

The qualitative approach presented in section 3 was developed to evaluate perceived audio quality during real MED situations while considering the perceived video quality and interaction between the teacher and student. The MED situations were real master classes in classical music taught over distance and connected via video conferencing systems and an IP network (Public Internet). That is, the teacher was at one location and the student at another. This section presents the results from the qualitative approach.

The study was conducted under non-experimental conditions. Measurements of latency between the location and audio and video were not possible due to inaccessibility to measurement equipment. Appendix 3 provides information about what equipment was used at each location.

4.1 Background to the study

The master classes were conducted in Oulu (Finland), Helsinki (Finland), Piteå (Sweden), Olos (Finland), and Rovaniemi (Finland) on several occasions during the autumn and winter of 2009 and spring and autumn of 2010. The instruments/ensembles used in these master classes were violin, French horn, cello, and string quartet. Singing was also a part of some classes.

4.2 Method used for the evaluation

The steps used for evaluation of the master classes were as follows:

1. Questionnaire 1 (Brief);
2. Questionnaire 2 (Extended);
3. Ad-hoc analysis of the questionnaires;
4. Interviews; and
5. Analysis of the interviews (meaning condensation).
Both questionnaires were completed and the interviews were conducted after each lesson so as not to interrupt the performance of the teachers and students. The time between distributing the questionnaires and conducting the interviews after each master class was no longer than one hour.

During the first master class, Questionnaire 1 (three open-ended questions) was distributed to obtain a general overview of the participants’ experiences: their evaluation of the overall sound quality and their evaluation of their overall distance learning experience. A total of six questionnaires were collected. From this information, a second questionnaire was designed and distributed to collect more detailed descriptions from the participants. The questionnaires were distributed to all participants (teachers, students, and observers participating in the master classes). A total of 22 questionnaires were collected over a period of three master classes.

After the distribution of the questionnaires, a total of six interviews were conducted: two interviews with students participating in the master classes and this master class study and four interviews with two master class teachers. (One of the teachers was interviewed three of the four times in order to collect additional information from each session.) The second teacher was interviewed via e-mail and several follow-up e-mails were sent to encourage further reflection on topics deemed interesting by the interviewer thus the same procedure as in a regular interview was used.

4.3 Analysis
One person conducted the analyses of the questionnaire data and interview data, using traditional data analysis methods. The analyses were carried out using the methods presented in section 3.4.

4.4 Results from the analysis

4.4.1 Questionnaire analysis (ad-hoc analysis)
The first part of the analysis, finding major trends, gave four major themes/trends. The sound quality-related trend contained statements on the perception of the sound. Several answers stated that the instruction and music examples played from the teacher/student were perceived as good and clear:

“Good natural sound”.

“Sound was very clear; there were no problems understanding what the teacher said”.

The sound and video quality-related trend included the perception of an asynchronicity between audio and video:

“The delay could be shorter”.

“The sound and the picture should be in the same time”.

The teaching-related trend showed that the participants thought that the distance master classes offered more opportunities for participating in master classes and that they offered the
opportunity to play in front of different teachers. This trend also meant travel was not required:

“The teaching was excellent”.

“You don’t have to travel far away to get lessons”.

The communication-related trend contained the perception of having a hard time communicating with the student or the teacher and playing together with the teacher:

“It is more difficult to communicate with the teacher”.

“There should be a clear signal to the students so that they would know when to stop playing”.

The second part of the ad-hoc analysis – meaning categorisation analysis (Table 1) – shows the total number of responses of the categories. Table 1 also shows the distribution of the received answers: the results show that the most frequently occurring responses are positive and are related to teaching. The second largest quantity of responses is positive and related to sound quality. The third largest quantity of responses does not contain any interpreted attitudes of positive, negative, both, or general statements (blank). These qualities are related to sound and video quality. The data also show that positive attitudes are most common. The second most common attitude is general statements without any positive/negative or both attitudes in the answers. The third most common attitude is negative and the fourth most common contains both positive and negative attitudes.

4.4.2 Interview analysis (meaning condensation)

The descriptive texts collected from each interview (step 5, Interview analysis) are summarized in this section into corresponding themes, i.e., making one descriptive text/description for each theme based on all interviews. This strategy is done to maintain the richness of the information in the collected data. A total of three themes were created.

Perceived audio and video quality

The teachers perceive that the video quality is good and works for distance master classes. The teachers could also imagine how the sound of each instrument sounded live, based on experience, even though the instruments did not sound natural during the master class. One teacher perceived the sound quality as “metallic” and “boring”, but with significant direct sound, no room sound, and good dynamics. The same teacher compared the sound to a high fidelity MP3. One teacher perceived a delay between audio and video, with the audio leading. With respect to sound quality, one teacher found it hard to distinguish between the system’s limitations and the student’s limitations although the teacher could distinguish this difference in a later master class. One teacher could easily see the student’s playing technique, but could not evaluate it because the sound was difficult to hear.

Both students compared the video quality as equal to or better than a YouTube clip. One student could only see the main features but not the contours or the proportions of the image. One student perceived the audio quality as “far away, distant, and a little muddy” and compared the audio to a YouTube clip. In addition, the same student found it hard to understand the teacher’s voice. One student compared the sound quality to a MP3 coded sound, but worse than a movie although the student could distinguish between a normal
spoken voice and a softly spoken voice. One student perceived the delay between audio and video as strange.

Perceived problems and possibilities
The delay between audio and video and the locations are perceived as a problem that makes it hard to communicate with the teacher/student. Small details in the music disappeared. Not knowing what is sent to the other locations, both related to video and sound quality, is perceived as a problem. During a distance master class, it is also hard to perceive the playing technique used by the students and how the students control their muscles. Distance learning is also problematic for music classes because it is difficult for teachers to evaluate their students’ playing techniques and muscle control.

Several topics – controlling tempo, intonation, articulation, and phrasing – were adequately dealt with during the classes. According to one teacher, if the delays between audio and video were short and the teachers were aware of it, the delay could be less problematic. Both teachers perceived the technology in an overall positive light. One student stated that physical presence, i.e., meeting in real life, is preferred over distance master classes, but distance master classes provide the opportunity to contact people who are based far away and to obtain new input and comments from other teachers. One teacher offered a similar evaluation. One teacher stated that there are indirect benefits for the master classes: saves time, travel, and money.

Perceived differences and similarities between regular and distance master classes
The perceived similarities were meeting with the student personally and discussions with the student. In addition, from a pedagogic point of view, the distant learning master class was similar to a regular master class. Two differences were identified: correcting playing technique was difficult and verbal explanations rather than hands-on demonstrations were required to explain new positions to the student, a situation that required more time. In addition, creating a relationship with a teacher/student during distance master classes was more difficult than during a regular master class. One teacher also stated that a physical/personal contact established before starting long periods of distance lessons would be helpful.

5. Discussion

5.1 The results

Seen together, the ad-hoc analysis and the meaning condensation analysis provided an overview of what type of trends/themes exist in the data and how the perceived audio quality relates to the perceived video quality and to the interaction between the users.

Masum et al. found that teachers and students found the tools (system used) comfortable [2]. This finding coincides with our results; the subjects could complete the master classes without any major difficulties. Our results also indicate that the perceived audio and video quality in the system is not optimal but is sufficient for this type of music education. There are also problems with lack of synchronicity between video and audio and latency between the locations, but the subjects’ statements indicate that they can work around these problems and manage a MED situation successfully. Woszczyk et al. also report similar results [4].
As a side note, the results from our study also align with previous research when it comes to teaching. To conduct successful teaching with the use of video conferencing systems, the teacher needs to adapt the content to handle the pedagogical situation [1]. In addition, it is cost-effective to bring teachers and teaching experiences to a large population of students [1]. Clearly, videoconference systems also allow teachers to train and teach in places other than their home location [2]. All these results coincide with our results.

The subjects’ initial attitude towards using such a system could be a bias that affected the results. That is, a student may have entered the study with a preconceived idea about distance learning. Such preconceived attitudes need to be considered when dealing with subjective responses. Before the study, based on their experience of sound quality, the authors expected the subjects to be negative about the sound quality; however, the results did not indicate this. Another possible bias, which may be connected to the positive attitudes collected, is that all the participants volunteered to participate in the study, showing they had an initial interest in MED; that is, the participants were self-selected on some level.

5.2 The methodology

Using qualitative methods including analysis shows a potential for arriving at a set of data that can be usable when evaluating several perceived qualities in one system. This method, with post questionnaires and post interviews, allows the participants to complete their task without interruptions and encourages the subjects to use their own words to describe their experience. This approach helps create a broad picture of what is happening in the study. This method also allows for a simultaneous description of what is perceived and what is affected even if there are more parameters affecting the perceived total quality than just the audio quality. Comparing this method to other subjective assessment methods of audio quality proposed by other studies [7], one could not say a particular method is better than another method because they collect different sets of data; however, some subjective assessment methods [5,6,7,8,9,11] have predetermined verbal descriptors and factors that aim at a defined part of the perceived quality. The approach in this paper enables the subjects to reflect on what they have perceived with few restrictions. Thus, the information can shed light on broader aspects of the perceived quality. This broad approach makes it possible to discover unexpected and possibly important factors, related either to audio quality or to other qualities, that affect how a specific situation or implementation is perceived. Hence, factors outside the audio domain may also be considered.

On the topic of analysis the data presented in Table 1, do not say anything in detail about the content of the responses; the data only show the distribution of answers, which demonstrates that simple quantitative observations can also be made using this method.

One person, as stated in the analysis section, analysed the questionnaires and interviews. Adding a second or several other researchers’ analyses and comparing them to each other could enhance the results further, as this allows checking the analyses for co-researcher agreement.

By using open-ended questions in a questionnaire, a variety of answers can be elicited. This strategy can be an advantage if a participant gives an answer that sheds light on a new area that the researcher has overlooked, or it can be fruitless if the participants “don’t know” or do not even answer the question. Open-ended questions used during an interview can encourage the interviewee to answer freely while still allowing the interviewer to guide the interviewee
Nyberg, Berg

A qualitative approach...

into interesting topics if they arise. In the latter situation, the interviewer needs to be very responsive to follow-up possibilities and needs to understand what to ignore. Such a strategy, of course, is a source of bias that has to be considered carefully. Researchers using questionnaires and conducting several interviews during an evaluation process must be aware that the data size quickly becomes large and is hard to manage and time-consuming to analyse.

In the meaning condensation of the interviews, each unit of meaning was categorized into themes. This division made the answers more clear and provided a better overview and a better understanding of each interviewed subject’s perception.

As all data collection was done after the completion of the event, this collection method relies on the subject’s memory. When answering questions after an event rather than during it, some shift in the perceived sensations may occur. This shift can be a disadvantage compared to other more direct methods used for rating and assessing. Small differences between sessions may not be captured, as they would be harder to detect when one experience is compared to another with a time gap in between.

Because the study was conducted during a real MED situation, some other limitations did arise. One limitation of the methodology is that only one person conducted the coding. However, when using a phenomenological or hermeneutical approach for the analysis, the researcher has to be aware of his/her own preconceptions and prejudices and exclude them during the analysis. The author performing the analysis was aware of this possible bias.

As can be seen in section 4, appendix 3, the equipment was changed occasionally between different locations and sessions. This can, of course, result in a bias, but the primary focus was not to link a set of experiences to particular equipment, but to evaluate the perceived audio and video quality in an ecological valid situation during live distance master classes. Another restriction was the unavailability of measuring the delay between the locations and between audio and video. This could have shed some light on when delay was present and when the delay was not present.

As indicated, the approach used in the current study yields different information compared to most of the previously used methods that evaluate audio quality. Hence, this method cannot be used interchangeably with existing methods to obtain the same type of data. However, by adding information that is not available from other methods, this approach will increase the knowledge of the subject’s experience. The results may also be used in an exploratory way as a means of observing what subjects perceive as noticeable, which in its turn can be used to develop evaluation scales used in existing methods. The rich verbal data resulting from a qualitative approach may provide a more holistic representation of an audio event, improving our understanding of how the event is experienced.

6. Acknowledgements

This work is part of the Vi R Music project and was partially financed by the European Union program Interreg IV A Nord. The authors would like to thank all the participants, students, and teachers who participated in the surveys and interviews. A thanks also goes to colleagues, technicians, and staff members who made this work possible. The authors also want to express a special thanks to Dr. Anders Persson, Jonas Ekeroort, and Jon Allan for their valuable comments on the drafts.
Nyberg, Berg

A qualitative approach…

References


8. BS.1116-1; Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems; ITU-R Recommendations (1994-1997)


11. BT.500-12; Methodology for the subjective assessment of the quality of television pictures, ITU-R Recommendations (09/2009)

12. Seshadrianthan, K; Soundararajan, R; Bovik, A, C.; Cormack L, K; Study of Subjective and Objective Quality Assessment of Video; IEEE transactions on image processing, Vol 19, No 6; June 2010


17. Jumisko-Pyykkö, S; Reiter, U; Weigel, Chr, Produced Quality is Not Perceived Quality - A Qualitative Approach to Overall Audiovisual Quality; *3DTV Conference*, May 7-9, 2007.


21. Häkkinen, Jukka; Alatonen, Viljakaisa; Schrader, Martin; Nyman, Göte; Lehtonen, Miikka; Takatalo, Jari; Qualitative analysis of mediated communication experience; *Quality of Multimedia Experience (QoMEX)*, Second International workshop, June 2010.


30. Nacke, Lennart E.; Grimshaw, Mark N.; Lindley, Craig A.; More then a feeling: Measurement of sonic user experience and psychophysiology in a first-person shooter game; *Interacting with computers* 22, pp. 336-343; 2010

31. Hsu, Shang Hwa; Wen, Ming-Hui; Wu, Muh-Cherng; Exploring user experiences as predictors of MMORPG addiction; *Computers & Education* 53, pp. 990-999, 2009


Figure 1. The steps and the results for each step used in the evaluation method.
Table 1. The number of responses in each category from the meaning categorisation analysis from the extensive questionnaire survey. The bold numbers show the categories attaining the largest number of statements in total and for each attitude.

<table>
<thead>
<tr>
<th>Descriptive features</th>
<th>Attitudinal features</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Tch</td>
<td>35</td>
<td>4</td>
</tr>
<tr>
<td>Sqr</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>Sqr&amp;Vqr</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Diverse</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Com</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Vqr</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Tec</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Tch&amp;Com</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>69</td>
<td>26</td>
</tr>
</tbody>
</table>
Appendix 1.

Questionnaire: Questions in bold are “yes” and “no” questions and the subject could only give a “yes” or “no” answer. The remaining questions are presented as open-ended questions in the questionnaire.

<table>
<thead>
<tr>
<th></th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Did you change your teaching methods on this occasion?</td>
</tr>
<tr>
<td>2</td>
<td>In what way did it change?</td>
</tr>
<tr>
<td>3</td>
<td>In what way were you able to interact, by speaking and singing/playing, with the teacher/student?</td>
</tr>
<tr>
<td>4</td>
<td><strong>Can you perceive the teacher’s instructions clearly?</strong></td>
</tr>
<tr>
<td>5</td>
<td>If “Yes”, describe in your own words the perceived sound.</td>
</tr>
<tr>
<td>6</td>
<td>If “No”, describe in your own words the perceived sound.</td>
</tr>
<tr>
<td>7</td>
<td><strong>Can you perceive the music examples from the teacher clearly?</strong></td>
</tr>
<tr>
<td>8</td>
<td>If “Yes”, describe in your own words the perceived sound.</td>
</tr>
<tr>
<td>9</td>
<td>If “No”, describe in your own words the perceived sound.</td>
</tr>
<tr>
<td>10</td>
<td>In what way did the communication between you and the teacher/student work?</td>
</tr>
<tr>
<td>11</td>
<td>What could further enhance the lesson?</td>
</tr>
<tr>
<td>12</td>
<td>Are there any positive aspects with the use of distance learning?</td>
</tr>
<tr>
<td>13</td>
<td>Are there any negative aspects with the use of distance learning?</td>
</tr>
<tr>
<td>14</td>
<td>What could be improved with the technology based on this lesson, according to your opinion?</td>
</tr>
<tr>
<td>15</td>
<td>What could be improved in the teaching, based on this lesson, according to your opinion?</td>
</tr>
</tbody>
</table>
Appendix 2.
Questions used in the interview guide. Translated from Swedish.

Start the interview with questions about name, age, etc.

- How was the lesson/master class?

- On the topic of user freedom
  - Did you feel limited in your practice?
    - Is it the technology?
    - Is it the distance between you and the teacher/student?
    - Is it the lack of presence by the teacher/student?
    - Is it the communication that poses problems?
  - If you feel free in your practice
    - What is it that makes you feel free in your practice?
    - Can you do the things that you want to do?
    - To what extent does the feeling of freedom exist?

- Could you complete the master class without being affected by the technology/system used?

- Did you perceive the technology as a hindrance or as a tool?

- What worked and what didn’t work during the master class?
  - Technology
    - Was it good?
    - Was it bad?
    - What could be improved?
      *Describe them*
  - Pedagogically
    - Was it good?
    - Was it bad?
    - What could be improved?
      *Describe them*

- How did you perceive the sound quality?
  - Can you compare it to a known format/media?

- How did you perceive the sound and video quality?

- If you exclude the video, how did you perceive the sound quality?

- If you exclude the sound, how did you perceive the video quality?

- Did you perceive any “delay/latency” between the sound and video?
  - If yes, which came first according to you?
  - Did this delay pose any problem for you?
Table with the equipment for each location.

<table>
<thead>
<tr>
<th>Location</th>
<th>Equipment</th>
</tr>
</thead>
<tbody>
<tr>
<td>All five locations</td>
<td>Tandberg MXP Edge 95 video conferencing systems, 50-52 inch LCD television screens</td>
</tr>
<tr>
<td>Piteå</td>
<td>2x Neumann KM184 microphones (occasionally a Microtech Gefell UMT 70s microphone)</td>
</tr>
<tr>
<td></td>
<td>2x Genelec 1030A speakers</td>
</tr>
<tr>
<td>Helsinki</td>
<td>2x Neumann TLM-103 microphones</td>
</tr>
<tr>
<td></td>
<td>2x Genelec 8030A speakers</td>
</tr>
<tr>
<td>Olos</td>
<td>2x Neumann TLM-103 microphones</td>
</tr>
<tr>
<td></td>
<td>2x Genelec 8030A speakers</td>
</tr>
<tr>
<td>Oulu</td>
<td>2x Neumann TLM-103 microphones</td>
</tr>
<tr>
<td></td>
<td>2x Genelec 8030A speakers</td>
</tr>
<tr>
<td>Rovaniemi</td>
<td>TV speakers</td>
</tr>
<tr>
<td></td>
<td>1x Clockaudio limited C600 microphone</td>
</tr>
</tbody>
</table>
Paper 2:  
Perceived audio quality of realistic FM and DAB+ radio broadcasting systems

Submitted to: Journal of the Audio Engineering Society
Perceived audio quality of realistic FM and DAB+ radio broadcasting systems

Jan Berg¹, Christofer Bustad², Lars Jonsson², Lars Mossberg² and Dan Nyberg¹

¹Luleå University of Technology
²Swedish Radio

Abstract
The perceived audio quality of a digital broadcasting system such as DAB+ depends on what type of coding and bit rates are applied. Due to bandwidth constraints, audio quality is prone to be in conflict with other service demands such as the number of channels and the transfer of ancillary data. Compared to DAB+, several other audio services have superior bit rates that challenge the audio quality of DAB+. An ideal level of quality is where no perceived difference between an original reference signal and its broadcasted counterpart is detected, sometimes referred to as perceptual transparency. This paper reviews and implements quality criteria used for investigating the quality and transparency of DAB+ codecs at higher bitrates and realistic FM and DAB+ systems including processors and low and high bit rate codecs commonly encountered in professional audio broadcasting. Two listening experiments using the ITU-R recommendations BS.1116 and BS.1534 (MUSHRA) and a closing interview were performed. In the MUSHRA test, a new type of anchor signal (one without low-pass filtering) was used. The results showed that the currently highest available subchannel bit rate for DAB+ (192 kbit/s) was insufficient for attaining perceptual transparency for critical items, but comparable to or in some instances better than a modern FM system. Extrapolation of data indicates that critical items may need even higher bit rates to be judged as being perceptually transparent. A linear relation between the MUSHRA and the BS.1116 scales was also suggested. From the interviews, auditory features important for the subjects’ assessment of quality were observed. This study concludes that when making decisions on broadcasting systems, it is important to have well-founded and clearly defined criteria for acceptable quality and/or transparency.

1 Introduction
Often broadcasters have heated and unresolved debates about the best bit rates to use when digitally broadcasting via Web radio and over DVB or digital radio airwaves. When designing a digital service, there are two contradicting targets that should fit within the decided investment level. One target is using high enough bit rates so perceived audio quality is deemed acceptable; the other target is the capacity to transmit as many channels as possible. The total cost of ownership of networks and transmitters often plays a dominant role for selection of bit rates, limiting the upper level of audio quality. Consequently, striking an appropriate balance between the number of channels and their audio quality is a delicate and crucial decision. In broadcasting, the relationship between perceived audio quality and bit rates is continuously being evaluated and discussed within and between radio and television companies and research bodies. In some cases, the selected bit rates are determined simply by just testing their impact on the public during a real broadcast; alternatively, bit rates are sometimes based on listening test results. The answers to questions about what
audio quality is desired and acceptable, how this relates to descriptors such as “good enough”, “good”, and “transparent” and what these actually correspond to in terms of bit rates are crucial for making an informed decision about the requirements of a broadcasting system.

Distribution or emission of audio over the airwaves tends to aim for minimal bit rates to cope with the ever-more crowded and expensive radio spectrum; however, earlier limitations for the maximum possible bit rates to be used in digital services via airwaves or the Internet to the home or to smartphones are gradually disappearing. A rapidly changing evolution can be seen towards higher possible and affordable bit rates. This change, of course, will allow broadcasters to aim for a higher audio quality, or, alternatively, to make more channels possible at a given audio quality level.

Due to bandwidth limitations, the general public has been forced to listen to quite low bit rates such as DAB in European countries and MP3 downloads used by portable devices. As these limitations become more and more obsolete, the question of whether audio quality levels are too low for the distribution and emission of DVB audio and digital radio should be addressed. A part of this discussion includes examining whether it is acceptable that this level of quality is lower than the quality of other popular consumer systems, such as CD, DVD, Blu-ray, and FM radio. In addition to physical media, downloadable high-quality counterparts\(^1\) in different formats (e.g., 5.1) exist as well as streaming applications such as iTunes\(^2\), Spotify\(^3\) and other formats that also provide listeners with higher bit rate audio.

For the assessment of the perceived audio quality of systems using perceptual coding, two predominant recommendations exist: ITU-R BS.1116 (ITU, 1997) and ITU-R BS.1534, often referred to as MUSHRA (ITU, 2003). These have been employed in numerous tests and have produced valuable results. The BS.1116 is designed to deal with small impairments, whereas MUSHRA is designed to assess intermediate quality. The border between these quality ranges, however, is not clearly defined. If the quality span of the tested systems is quite wide, the upper part of the quality space may call for the use of BS.1116, while the lower part would be more suited for a MUSHRA test. In conjunction with such considerations, informal discussions have focused on the relation between these two scales and what types of degradations subjects experience during such tests.

This paper investigates the perceived audio quality of FM and DAB+ systems at different bit rates and whether the quality provided is sufficient. This study has five objectives: i) to assess the higher bit rates in DAB+ as well as realistic FM and low and high bit rate DAB+ systems by means of different definitions of audio quality; ii) to compare these FM and DAB+ systems with one another; iii) to investigate what additional information about the perceived audio quality of

---

\(^1\) E.g. www.hdrtracks.com/, www.rdio.com, www.wimpmusic.se

\(^2\) www.apple.com/itunes/

\(^3\) www.spotify.com/
such systems can be elicited by conducting interviews after listening tests; iv) to make observations on the relation between MUSHRA and BS.1116 evaluation scales; and v) to estimate what bit rates are likely to be required for transparency. The term “realistic system” here refers to the chain of processors and codecs that would be commonly encountered in professional audio broadcasting. The systems contribute to the ecological validity of the results of the upcoming experiments.

2 Audio quality in broadcasting systems

A broadcasting system generally consists of circuits for contribution, distribution, and emission. These refer to different parts of the system: contribution – the network between production sites; distribution – the network delivering the programme to the transmitter; and emission – the RF signal.

When discussing audio quality, several possible quality criteria are used, especially the concept of “transparency”, which can be interpreted in several ways. One common use is “bit transparent”. Bit transparent refers to when a PCM sample passes through some apparatus unaffected with identical content on input and output bit-by-bit. A second definition might be called “perceptual transparency”. Perceptual transparency means people cannot perceive changes in audio quality when comparing audio to the original (before encoding or processing). This means that in a listening test the score of a perceptually transparent item should not show any statistically significant difference from the reference signal. A third notion, “acceptable” broadcast audio quality, requires a better score for all items than one grade down on the 5-grade evaluation scale used in ITU-R BS.1116, i.e., a Subjective Difference Grade (SDG) > -1. This limit is the lowest allowed result of a listening test for production or contribution circuits. In EBU Tech 3339, this interpretation is described in the following way:

If cascaded codec chains are to be considered for broadcast use then the quality criterion should be that none of material should produce an average diff-grade worse than -1.0 (“perceptible but not annoying”). If all the tested items score better than -1.0, then we can consider the chain’s performance to be acceptable. However, if the average over all items is better than –1.0, but some test items score worse than this, then we must be wary of using such a chain. If the average over all items is worse than -1.0 then the chain should not be considered acceptable for broadcast use. (EBU, 2010)

As this quotation suggests, even when a specific codec shows an overall acceptable performance averaged across items, certain single items may still show an unacceptable quality, a concern that any analysis should consider. Similarly, EBU BPN 019 (EBU, 1999) uses the phrase “indistinguishable quality” and EBU BPN 094 (EBU, 2009) uses the phrase “broadcast quality” to define acceptable quality. The latter notion seems to be implicitly defined as excellent quality or a score > 80 on the MUSRHA scale in EBU BPN 094, Section 9.1.
In addition, there is a possible fourth audio quality criterion, “FM quality”. As FM has been the predominant way of audio broadcasting, it is a de-facto reference point, or a sort of anchor that later systems can and will be compared with by listeners. Consequently, a comparison between DAB+ and FM is of interest. Such a comparison would not be straightforward as the systems under test will be susceptible to different forms of quality degradation under real-life conditions. However, if conditions are specified, a comparison is possible.

Audio over IP contribution circuits require higher bit rates and thus bandwidths than those used in emission for distribution systems. This is a consequence of the fact that a significant number of steps of re-encoding must be made in the contribution chain, e.g., in editing or other manipulation of the programme. As a result, cascading artefacts are added to the final programme before distribution. To avoid such artefacts, the bit rate has to be increased (Geiger, o.a., 2007) (EBU, 2005) (EBU, 2007) (EBU, 2010). Although this fact has been reported, it is not widely known so it has not influenced the roll out of bit rate compressed audio. More discussion, research, and listening tests on the cascading performance of, for instance, AAC are needed to understand the phenomenon even better.

The cost for IP networks used for high bit rates is slowly decreasing. With lower prices for higher bandwidths coming, this actually allows the use of sufficiently high bit rates to avoid the cascading artefacts. Figure 1 shows the evolution over time of available bandwidths (Gross, 2011). Today, there is no reason to keep the very low bit rates that started to be used when audio over IP contribution systems were introduced just after 2000. There is a continuous increase in the actual bit rates over all IP networks at a more or less unchanged price.

![Figure 1. Evolution of available bandwidth for backplane (i.e., in switches), Ethernet hardware, and the Internet (Gross, 2011).](image)
Examples of different systems and the associated codecs and typical bit rates are found in Table 1. Clearly, today typical DVB and consumer systems use higher bit rates than digital radio broadcasting systems.

Table 1. Examples of codecs and typical bit rates used in current consumer systems and in DAB and DAB+.

<table>
<thead>
<tr>
<th>System</th>
<th>Examples of codecs</th>
<th>Typically used bit rates (kbit/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAB+</td>
<td>AAC/HE-AAC</td>
<td>32-128</td>
</tr>
<tr>
<td>DAB</td>
<td>MPEG-1/2 Layer II</td>
<td>64-192</td>
</tr>
<tr>
<td>DVB</td>
<td>MPEG Layer II</td>
<td>192-256 (stereo)**</td>
</tr>
<tr>
<td></td>
<td>Dolby Digital</td>
<td>448 (stereo + multichannel)</td>
</tr>
<tr>
<td>Blu-ray discs</td>
<td>PCM*/Lossless coding</td>
<td>≥6 Mbit/s (stereo + multichannel)</td>
</tr>
<tr>
<td>DVD</td>
<td>PCM*/DTS/Dolby Digital</td>
<td>2304 (stereo), 640-1500 (multichannel)</td>
</tr>
<tr>
<td>Online music catalogues</td>
<td>FLAC*</td>
<td>≥800 (stereo)</td>
</tr>
<tr>
<td>Web streaming</td>
<td>MPEG-1 Layer II/Windows Media Audio/AAC</td>
<td>32-320, 128 typical</td>
</tr>
<tr>
<td>iTunes</td>
<td>AAC</td>
<td>128-256</td>
</tr>
<tr>
<td>Spotify</td>
<td>Ogg Vorbis</td>
<td>96-320</td>
</tr>
<tr>
<td>Wimp</td>
<td>AAC/HE-AAC</td>
<td>64-256</td>
</tr>
</tbody>
</table>

* PCM and FLAC mean that no lossy data compression has been used.
** Bit rates used in Sweden.

A more detailed look at typical bit rates in DAB+ can be found in Figure 2.

![Figure 2. The sum of occurrences of different bit rates in DAB+ in the countries Australia, Germany and Switzerland in May 2012 [http://Wohnort.org/DAB/index.html].](image)

Clearly, several possible definitions of audio quality of a system exist and a wide span of bit rates, many of them lower than what is employed by other audio applications, are used for DAB+. The total signal path of digital broadcasting
includes multiple audio coding and decoding points where signal degradations occur. To handle these issues for future audio broadcasting, it will be important to understand which quality criteria are suitable and how these should be interpreted and implemented.

3 Listening tests – coding processes

In the following experiments, a number of coding processes possible for broadcasting purposes will be investigated for their perceived audio quality. In this section, the different processes will be described. Two test methodologies were used in these experiments, which were divided into Test 1 and Test 2. Test 1 used BS.1116 (ITU, 1997) and Test 2 used MUSHRA (ITU, 2003). Further details on the experiments can be found in Section 4. Loudness levels in LUFS refer to recommendation EBU R128 (EBU, 2010).

3.1 Test 1

One objective of this test was to assess the higher bit rates in DAB+ by means of different definitions of audio quality. Another objective was to make observations on the relation between the scales of the MUSHRA and BS.1116 evaluation methods by comparing results from this listening test, which used the BS.1116 method with results from the EBU D/DABA project report BPN 094 (EBU, 2009) for which the MUSHRA method was used. For this reason, the same encoder and decoder and the same settings used for the test described in BPN 094 were also used for this test.

All audio selected for the test was encoded into the specific type of HE AAC used in DAB+ using a command line software encoder called “testenc” version 1.2.0 (built July 16, 2007) from Dolby and then decoded using the Dolby command line decoder “testdec” version 1.0.0 (build 31 May 2007).

In DAB+, the bit rate for a subchannel is not only used for audio data but also for programme associated data (PAD), error correction, etc. The bit rate available for audio data is about 88-91% of the subchannel bit rate (due to error correction) minus the bit rate used for PAD (ETSI, 2010). The highest subchannel bit rate allowed in DAB+ is 192 kbit/s and to keep the number of different bit rates at a level that would avoid a too lengthy test and possible listener fatigue, four bit rates spaced 32 kbit/s apart were chosen. Consequently, the subchannel bit rates used by the encoder were 96 kbit/s, 128 kbit/s, 160 kbit/s, and 192 kbit/s. Spectral band replication (SBR) was only used for the lowest bit rate.

The bit rates for PAD were 1.33 kbit/s for the lowest bit rate with SBR and 2.53 kbit/s for the three higher bit rates without SBR. These values were chosen to be the same as in the listening test reported in EBU BPN 094 (EBU, 2009). There are, however, a number of services for which higher PAD bit rates probably would be selected and the maximum bit rate for PAD can be set as high as 79 kbit/s for 48 kHz sample rate when SBR is disabled. If a higher PAD bit rate is selected and if the subchannel bit rate remains the same, the audio quality will decrease. The audio bit rates including PAD were 87.2 kbit/s, 115.8 kbit/s, 145.1
kbit/s, and 174.5 kbit/s. For more details of the PAD insertion, see Section 7 in EBU BPN 094. The sampling frequency was always set to 48 kHz and all audio was encoded in stereo. See Table 2 for names of the settings and an overview of the bit rates.

Table 2. Codec settings in Test 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Subchannel bit rate [kbit/s]</th>
<th>Audio bit rate (including PAD) [kbit/s]</th>
<th>PAD bit rate [kbit/s]</th>
<th>Spectral Band Replication (SBR)</th>
<th>Parametric Stereo (PS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAB+ 96 (SBR)</td>
<td>96</td>
<td>87.2</td>
<td>1.33</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 128</td>
<td>128</td>
<td>115.8</td>
<td>2.53</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 160</td>
<td>160</td>
<td>145.1</td>
<td>2.53</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 192</td>
<td>192</td>
<td>174.5</td>
<td>2.53</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

One excerpt in this listening test (Speech (pan)) was encoded and decoded in two versions. The first version was the original speech signal and the second version was the same audio sent through an MPEG-1 Audio Layer II encoder at 384 kbit/s, stereo, 48 kHz, and a decoder five times. The encoder and decoder was a special version of Awave Audio by FMJ Software called Awave SR version 2.3, which internally uses TooLame. The audio bandwidth of this encoder at this bit rate exceeded 20 kHz. The production at broadcasting companies often includes a number of cascades, e.g., for Swedish Radio this is valid for MPEG-1 Audio Layer II. Hence, to investigate its influence on the audio quality the cascade was included.

Each item to be used in the listening test was finally adjusted to -23 LUFS and was also carefully synchronized with other items originating from the same excerpt. See Figure 3 for a simplified graph of Test 1.

![Figure 3. Block diagram of Test 1.](image)

### 3.2 Test 2

Test 2 assessed realistic FM and DAB+ systems using different definitions of audio quality and compared FM and DAB+ systems. As this is common practice in today’s broadcasting, FM and DAB+ in this experiment both use one band and multiband audio processing to reduce unintentional level differences, to adapt the dynamics of the audio content for the listening environments common to listeners, to reduce the frequency response differences between different audio content, and, in some cases, to create a certain “sound” for a particular channel. In FM, one band (sometimes referred to as wide-band) and multiband processing is also usually closely adapted to the pre-emphasis limiting necessary for the FM stereo system. To create realistic broadcast systems, it was decided to include one band and multiband audio processing in the DAB+ and FM systems to be tested.
Since different audio content requires different types of audio processing, two processing categories were created. One category, P2, corresponded to the radio channel Swedish Radio P2, which broadcasts classical and contemporary music as well as jazz and folk music. The other category, P3, corresponded to the radio channel Swedish Radio P3, which broadcasts pop music, news, and cultural and social programmes. The processing for the P2 category was decided to include only one band audio processing and no multiband audio processing and the processing for the P3 category was decided to include one band processing followed by multiband audio processing.

To compare the audio quality of FM and DAB+ on the same terms, it was important for the audio processing to be the same for both systems, with the exception of the pre-emphasis limiting for FM. To ensure this consistency, audio content that later might be selected for the listening test was sorted into the two categories, P2 and P3. The total amount of audio in the two categories was about seven hours and was later reduced to a few excerpts. The loudness of each audio item in both categories was first normalized to -23 LUFS. Then audio in the P2 category was processed using the advanced one band audio processor Factum Cadenza⁴ in such a way that the resulting loudness for everything in the category as a whole reached -23 LUFS. Using this strategy meant that the overall loudness corresponded to the recommended loudness according to EBU Tech 3344 (EBU, 2011). The preset used, Preset P2, was a slight modification of the preset currently used by the Swedish Radio P2 channel during evenings. Preset P2 performed a minimum of short-term limiting and mainly raised the level of soft passages in the audio.

The audio in the P3 category was processed using the same one band audio processor followed by the software multiband audio processor Breakaway⁵. The preset used for the one band processor in this category, Preset P3, was a slight modification of the preset currently used by the Swedish Radio P2 channel during daytime. This preset also did a minimum of short-term limiting and mainly raised the level of soft passages in the audio, but more so than the Preset P2. The multiband processor used a gentle preset that did some dynamic compression with a moderate compression ratio in six frequency bands. The drive into its final limiter was set in such a way that the resulting loudness for everything in the category as a whole reached -23 LUFS.

The processed audio for the P2 and the P3 categories was fed into a DAB+ encoder and also into two different FM systems in which the broadcast processors only did pre-emphasis clipping and no other audio processing. The drives into the pre-emphasis clippers were set so that the MPX Power at some point reached but never exceeded the MPX Power limit defined in ITU-R BS.412 (ITU, 1998). In this way, all the systems used audio processing that is realistic and normally occurring in broadcasting. The reference signals were extracted

---

⁴ www.factum.se
⁵ www.claessonedwards.com
from the output of the audio processors (multiband/one band). See Figure 4 for a block diagram of Test 2.

![Block diagram of Test 2](image)

**Figure 4. Block diagram of Test 2.**

### 3.2.1 DAB+

Given the present time restrictions, only a subset of the seven hours of processed audio was fed into an actual DAB+ encoder. All seven hours of audio were encoded into HE AAC using a command line encoder from Dolby and decoded back to PCM. This type of HE AAC, however, was not exactly the one used in the DAB+ system, but should approximate the correct type well. The bit rates used by the command line encoder were set to the actual audio bit rates used by the DAB+ encoder, i.e., bits used in DAB+ for error correction and programme associated data (PAD) were taken into account.

The selected excerpts for the listening test were fed into the DAB+ encoder Factum MAP250E and the audio from the monitor output was recorded. The audio was also sent over the air simultaneously through a DAB+ multiplex and a transmitter placed in the Nacka region of Stockholm, and the audio from a DAB+ receiver was recorded. As no differences could be found between the audio from the monitor output and the receiver, the audio from the monitor output was chosen for the listening test.

The subchannel bit rates used by the DAB+ encoder were 48 kbit/s, 64 kbit/s, 96 kbit/s, 128 kbit/s, 160 kbit/s, and 192 kbit/s. Spectral band replication (SBR) was only used for the three lower bit rates, and parametric stereo (PS) was only used for the lowest bit rate. The bit rates for programme-associated data (PAD) were 1.3 kbit/s for the three lower subchannel bit rates and 2.7 kbit/s for the three higher subchannel bit rates. These values were chosen to be similar to those used in EBU BPN 094 (EBU, 2009) and in Test 1 in this paper. The audio bit rates including PAD were 43.2 kbit/s, 57.9 kbit/s, 87.2 kbit/s, 115.8 kbit/s, 145.1 kbit/s, and 174.5 kbit/s. The sample rate was always 48 kHz and all audio was encoded in stereo.
Each item to be used in the listening test was finally adjusted to -29 LUFS, not -23 LUFS as in Test 1. This adjustment prevented overshoots resulting from the audio codec from being clipped. In Test 2, one of the ten excerpts was a 22 dB attenuated copy of one of the other nine excerpts. Some of the items resulting from the attenuated version would be clipped in the normalisation if they were normalised to -23 LUFS as in Test 1. Hence the loudness level -29 LUFS was chosen for Test 2.

Some items resulting from the non-attenuated version of the excerpt, however, were clipped in the HE AAC encoding or decoding. It is not known if the clipping occurred in the encoding and/or the decoding, since the effects were observed in the decoded audio. This clipping occurred at the bit rates 48 kbit/s, 64 kbit/s, and 96 kbit/s at some transients occurring in the excerpt. Since the audio level at the input to the encoder corresponds to the normalisation standard EBU R128, which in its turn corresponds to real broadcasting conditions, these items were kept and used in the listening test. Audio coding has been shown to produce overshoots up to 5.3 dB (Nielsen & Lundh, 2003). So far, unpublished research by one of the authors of this paper has shown overshoots up to 8 dB for 64 kbit/s. See Table 3 for DAB+ codec settings and an overview of the bit rates.

<table>
<thead>
<tr>
<th>Name</th>
<th>Subchannel bit rate [kbit/s]</th>
<th>Audio bit rate (including PAD) [kbit/s]</th>
<th>PAD bit rate [kbit/s]</th>
<th>Spectral Band Replication (SBR)</th>
<th>Parametric Stereo (PS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAB+ 48 (SBR, PS)</td>
<td>48</td>
<td>43.2</td>
<td>1.3</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DAB+ 64 (SBR)</td>
<td>64</td>
<td>57.9</td>
<td>1.3</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 96 (SBR)</td>
<td>96</td>
<td>87.2</td>
<td>1.3</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 128</td>
<td>128</td>
<td>115.8</td>
<td>2.7</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 160</td>
<td>160</td>
<td>145.1</td>
<td>2.7</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 192</td>
<td>192</td>
<td>174.5</td>
<td>2.7</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

### 3.2.2 FM

The same processed audio that was fed into the HE AAC encoder and the DAB+ encoder was also fed into two FM systems. FM audio quality can mean many different things depending on the equipment used in the signal chain and in particular on the broadcast processor and how it is set to process the audio. For this reason, two FM systems were included in this listening test.

The first FM system, FM 1, consisted of a prototype of the audio processor Omnia 96 with 16 bit PCM input and output. Only its psychoacoustic multiplex clipper and a phase scrambler were active and no other processing was done in the processor. The phase scrambler could also be turned off. See section 4.2.2 for details on when the phase scrambler was in use. To attain an extended audio frequency range of about 17.5 kHz in this prototype, single sideband mode was

---

6 omniaaudio.com
active for the difference signals (S=L-R) in the FM multiplex. The audio processor also added an RDS signal with 6 kHz deviation.

In theory, an ideal FM modulation, demodulation, and reception would not affect the perceived audio quality. Hence, to get as close as possible to an ideal FM transmitter and receiver by removing the RF path, the multiplex output from the processor was saved directly to audio files at f_s = 192 kHz and decoded from multiplex to left/right audio using a command line utility⁷. To set the drive into the psychoacoustic multiplex clipper so that the multiplex power at some point reached but never exceeded the MPX Power limit defined in BS.412, a command line utility was used that calculated the maximum of a moving average of the multiplex power. See Figure 5 for a simplified graph of FM 1.

![Figure 5. Block diagram of FM 1.](image)

The second FM system, FM 2, is of the same type as the most common FM system in Swedish Radio’s FM network. This system’s signal path started with the same audio processor as in FM 1 that here used a psychoacoustic left/right clipper and a phase scrambler and not its psychoacoustic multiplex clipper or any other processing. In this audio system, the phase scrambler could also be turned off. See section 4.2.2 for details on when the phase scrambler was in use. De-emphasized left/right audio, limited to a 15 kHz bandwidth, and an RDS signal were fed into a Profline DMM stereo encoder with pre-emphasis enabled, and an analogue multiplex signal from the stereo encoder was fed into a BW Broadcast TX5 FM transmitter. A 30 dB RF attenuator then reduced the output level from the transmitter into an Audemat FM-MC4 receiver (upgraded in software from FM-MC3). The multiplex output from this receiver was recorded on a SoundDevices 722 unit at 192 kHz mono 24 bits. Then the command line utility also used in FM 1 decoded the multiplex audio to left/right audio.

To set the drive into the psychoacoustic left/right clipper so that the multiplex power at some point reached but never exceeded the MPX Power limit defined in BS.412, the same command line utility used in FM 1 was used on multiplexed audio before de-multiplexing and de-emphasis. The audio sent to the stereo encoder was left/right de-emphasized audio to imitate the current FM transmission chain for Swedish Radio. The audio was prepared in advance and played back on a SoundDevices 722 unit. The RDS signal sent to the stereo encoder was a band-pass filtered recording of an RDS signal played back on a SoundDevices 722 unit at 192 kHz sample rate. The deviation of the FM signal was set to 75 kHz, the RDS level was set to 3 kHz, and the signal strength into the receiver was set to 1 mV, all using measurements in the Audemat receiver. See Figure 6 for a simplified graph of FM 2.

⁷ Provided by Leif Claesson
Figure 6. Block diagram of FM 2.

Each item to be used in the listening test from the two FM systems was finally adjusted to -29 LUFS to match the loudness level of the DAB+ systems. Each audio item was also carefully synchronized with other items originating from the same excerpt.

3.2.3 MPEG Audio Layer II

It was discussed whether MPEG Audio Layer II as used in DAB should be included among the audio systems in this test. To avoid too many items to be graded by the subjects and since DAB using MPEG Audio Layer II will probably not be introduced in Sweden beyond trials, it was decided not to include MPEG Audio Layer II.

3.2.4 Anchor

According to the MUSHRA specification, at least one anchor should be included for post-screening purposes and to make sure that items in the test cover a large portion of the audio quality scale. The required anchor in the recommendation is a 3.5 kHz low-pass filtered version of the reference signal. This anchor has been used in many tests over the years, but the anchor has lately been questioned as the audio quality of modern audio codecs has improved significantly. Since most codecs tested today have a much wider bandwidth and exhibit completely different artefacts compared to a 3.5 kHz low-pass filtered version, this anchor always stands out even if a test item has severe coding artefacts. This is the reasoning given by the D/DABA group in the EBU report BPN 094 to explain why a new kind of anchor that does not employ band limitation was developed and used.
For the same reasons, the anchor developed by the D/DABA group in the MUSHRA test described in this paper was used. This new anchor is created by a simple algorithm that combines two types of signal impairment to reflect current typical coding artefacts. The first impairment is MDCT-based quantisation distortion and the second is stereo image distortion. This algorithm is described in detail in BPN 094 and in an open source implementation\textsuperscript{8}. The parameter settings used in this test were the same as used for the D/DABA tests (distortion set to 10 and fuzzy to 0.25) (EBU, 2009). The anchor items were adjusted to -29 LUFS to match the loudness from the other items and the anchor items were also carefully synchronized with other items originating from the same excerpt.

4 Listening tests – method
The listening tests were performed to investigate the perceived audio quality of items originating from excerpts processed through the different coding processes described in Section 3. The assessment method in Test 1 was the ITU-R BS.1116 recommendation (ITU, 1997); the assessment method in Test 2 was the ITU-R BS.1534 (MUSHRA) (ITU, 2003). In this section, further details on the listening tests will be provided.

BS.1116 was chosen to detect expected small differences between the systems tested in Test 1, whereas MUSHRA was used for the expected bigger quality differences due to the inclusion of the low bit rate systems in Test 2. Test 1 was always performed first by the subjects due to the expected smaller differences between codecs thus avoiding a harder task at the end when listener fatigue could influence the results.

4.1 Subjects
All the subjects (N=30) had experience listening critically to reproduced sound. They were recruited at two locations in Sweden. In Stockholm, the subjects (n=17; aged 21 to 62; mean=41 years; median=41) included personnel from Swedish Radio as well as independent listeners although most listeners were professional sound engineers. At the second location, the School of Music at Piteå, a campus at Luleå University of Technology, the subjects (n=13; aged 19 to 29; mean=23; median=21) all had audio technology education and were either students or alumni at the school. Six of the subjects had undergone hearing tests in their admission tests for their educational program. The remaining subjects were assumed to have regular hearing (none of them stated a known hearing loss before the test).

4.2 Programme material
The general criteria for selection of excerpts can be summarised as follows.

- The excerpts should span a broad range of different types of material and musical genres.

---

\textsuperscript{8} sourceforge.net/projects/anchor
• The excerpts should come from typical programme material.
• The excerpts should clearly reveal something about the performance of one or more of the systems and of differences between the systems.
• There must not be any bias towards or away from any particular system.
• The excerpts should both be material previously used in other listening tests (to compare with other tests) and material not previously used in other listening tests (to avoid the possibility that codecs might be tuned to the excerpts selected for the test).
• The excerpts should not be wearisome or too involving.

4.2.1 Test 1
Since the objectives of this test were to investigate the audio quality reached at the higher bit rates in DAB+ and to compare results from this listening test with results from EBU BPN 094 (EBU, 2009), four of the most critical excerpts used in BPN 094 were selected for this test. Including all the items from the previous test (which was performed according to the MUSHRA recommendation) would have made the length of the test impractical. Since the same encoder used for BPN 094 was used for this test, there was no particular need to seek new excerpts. The four excerpts selected are found in Table 4. The excerpt “Speech (pan) + 5x L2” is the excerpt “Speech (pan)” after five encodings and decodings of MPEG-1 Audio Layer II as described in Section 3.1. As previously mentioned, production often includes a number of cascades. Hence, one such excerpt was included.

<table>
<thead>
<tr>
<th>Name of excerpt</th>
<th>Name in BPN 094</th>
<th>Description</th>
<th>Length</th>
<th>Upper frequency limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applause w announcer</td>
<td>g_applause, and female announcer</td>
<td>Applause with female announcer.</td>
<td>16.5 s</td>
<td>18.3 kHz</td>
</tr>
<tr>
<td>Classical</td>
<td>f_brass, timpani and castanets</td>
<td>Brass, timpani and castanets, from Manuel de Falla’s ballet <em>El Sombrero de tres picos</em>. Taken from track 1 on the SACD HMC 801606 from Harmonia Mundi.</td>
<td>17.8 s</td>
<td>24 kHz</td>
</tr>
<tr>
<td>House</td>
<td>a_electro pop</td>
<td>Excerpt from the house/pop song “Love is Gone” by David Guetta.</td>
<td>20.1 s</td>
<td>22 kHz</td>
</tr>
<tr>
<td>Speech (pan)</td>
<td>b_female speech Swedish</td>
<td>Female speech from a Swedish newscast panned slightly to the right.</td>
<td>16.0 s</td>
<td>24 kHz</td>
</tr>
<tr>
<td>Speech (pan) + 5x L2</td>
<td>Not in BPN 094</td>
<td>Speech (pan) after five encodings and decodings of MPEG-1 Audio Layer II at 384 kbit/s.</td>
<td>16.0 s</td>
<td>20.5 kHz</td>
</tr>
</tbody>
</table>

4.2.2 Test 2
Since the objectives of Test 2 were to assess realistic FM and DAB+ systems using different definitions of audio quality and to compare FM and DAB+ systems, audio content was first collected that was thought to be critical for one or more of the systems in the test. The seven hours of audio from the different categories and musical genres that were sorted into the two audio processing categories denoted “P2” and “P3” (as described in Section 3.2) were reduced using the criteria in Section 4.2. The audio processing for each category was
performed and the processed audio was fed through the systems described in Section 3.2. The outputs from the different systems were then loudness aligned and synchronized with the original processed audio.

Using the training mode in the software STEP (see Sect. 4.4), it was possible to easily switch between the original processed audio and the outputs from the different systems, maintaining time synchronization. It was then noted which audio content was the most critical for the different systems. An overview of the ten selected excerpts can be found in Table 5.

<table>
<thead>
<tr>
<th>Name of excerpt</th>
<th>Preset</th>
<th>Description</th>
<th>Length</th>
<th>Upper frequency limit</th>
<th>Audio processing category</th>
<th>Phase scrambler in FM systems</th>
<th>Commonalities with excerpts in Table 4, with the exception of the audio processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appliance w announcer</td>
<td>P2</td>
<td>Appliance with female announcer.</td>
<td>16.4 s</td>
<td>18.4 kHz</td>
<td>P2</td>
<td>On</td>
<td>Same as &quot;Appliance w announcer&quot;</td>
</tr>
<tr>
<td>Classical</td>
<td>P2</td>
<td>Brass, timpani and castanets, from Manuel de Falla’s ballet El Sombrero de tres picos. Taken from track 1 on the SACD HMC 801666 from Harmonia Mundi.</td>
<td>17.8 s</td>
<td>24 kHz</td>
<td>P2</td>
<td>On</td>
<td>Same as &quot;Classical&quot;</td>
</tr>
<tr>
<td>Electronic</td>
<td>P2</td>
<td>Excerpt with strong transients from &quot;Vaihtovirta&quot; on the album Aaltopirsi by Pan Sonic.</td>
<td>25.5 s</td>
<td>22 kHz</td>
<td>P2</td>
<td>On</td>
<td></td>
</tr>
<tr>
<td>Electronic (att)</td>
<td>P2</td>
<td>Same as Electronic but attenuated by 22 dB. That is, the loudness is -51 LUFS.</td>
<td>25.5 s</td>
<td>22 kHz</td>
<td>P2</td>
<td>Off</td>
<td></td>
</tr>
<tr>
<td>House</td>
<td>P3</td>
<td>Excerpt from the house/pop song &quot;Love is Gone&quot; by David Guetta.</td>
<td>19.8 s</td>
<td>22 kHz</td>
<td>P3</td>
<td>On</td>
<td>Same as &quot;House&quot;</td>
</tr>
<tr>
<td>PopKent</td>
<td>P3</td>
<td>The intro to the song &quot;Dom Andra&quot; by Kent.</td>
<td>19.6 s</td>
<td>21 kHz</td>
<td>P3</td>
<td>On</td>
<td></td>
</tr>
<tr>
<td>PopRox</td>
<td>P3</td>
<td>A chorus in the song &quot;Fading Like A Flower&quot; by Roxette.</td>
<td>24.5 s</td>
<td>22 kHz</td>
<td>P3</td>
<td>On</td>
<td></td>
</tr>
<tr>
<td>SpeechL (no pan)</td>
<td>P3</td>
<td>Same as SpeechL (pan) above but without the panning.</td>
<td>26.4 s</td>
<td>24 kHz</td>
<td>P3</td>
<td>On</td>
<td></td>
</tr>
<tr>
<td>SpeechL (pan)</td>
<td>P3</td>
<td>Female speech (long) from a Swedish news cast panned slightly to the right.</td>
<td>26.4 s</td>
<td>24 kHz</td>
<td>P3</td>
<td>On</td>
<td>A longer version of &quot;Speech (pan)&quot;</td>
</tr>
<tr>
<td>World</td>
<td>P2</td>
<td>Excerpt with handclaps and not much other percussion from the song &quot;Migration&quot; on the album Introducing by Nt in Sawhney.</td>
<td>21.4 s</td>
<td>22 kHz</td>
<td>P2</td>
<td>On</td>
<td></td>
</tr>
</tbody>
</table>

Preset “P2” indicates audio processing category P2; i.e., that the audio processing was done using only an advanced one band audio processor that did a minimum of short term limiting and mainly low level compression. Preset “P3” indicates audio processing category P3; i.e., the audio processing was done using the same one band audio processor as for the P2 category, again doing a minimum of short term limiting but more low level compression, followed by a software multiband audio processor using a preset designed to limit its influence on the audio signal. The audio processing for both categories are described in more detail in section 3.2.
SpeechL, aside from the audio processing P3, is the same audio as Speech (pan) in Test 1 plus an additional 10.4 s of audio from the same newscast.

Electronic (att) is the same audio as Electronic but attenuated by 22 dB to make sure that no audio reached the clip level of the pre-emphasis clippers in the FM systems. This excerpt contains high-level transients at high frequencies, a condition that explains why the attenuation was chosen to be so high. After feeding this attenuated excerpt though the different systems, the level was restored through amplification of the resulting items to match the loudness of the other items.

The phase scrambler in the FM audio processor makes asymmetric audio waveforms more symmetric and thus less sensitive to pre-emphasis clipping. The reason for its use is that certain types of signals such as speech can have higher peak levels on the positive side of the waveform than on the negative side or vice versa, and this asymmetry can be reduced by the phase scrambler. When more symmetry is reached, the highest peak levels are usually reduced. Experience shows that the use of the phase scrambler may cause quality alterations. Hence, the phase scrambler was enabled for all excerpts except Electronic (att) as this was attenuated. The excerpt Electronic (att) was designed not to trigger the pre-emphasis clipper and thus the phase scrambler was not necessary and therefore turned off.

Below is a list of the selection of excerpts for Test 2:
- Five excerpts in the P2 audio processing category and five excerpts in the P3 category;
- Three excerpts containing modern mainstream music and four excerpts containing non-mainstream music;
- One excerpt containing applause;
- One excerpt containing panned speech;
- One excerpt containing non-panned speech;
- Two versions of an excerpt with strong transients with different levels to investigate how the pre-emphasis clipper influences the grading of the FM systems;
- Four excerpts that are, aside from the audio processing, the same as in Test 1 and in BPN 094; and
- Some excerpts containing a moderate amount high frequency energy and excerpts containing much high frequency energy.

All excerpts were edited to have smooth beginnings and ends and were calibrated to -29 LUFS so as to have the same loudness. See Section 3 for more information.

### 4.3 Systems under test – summary

The systems included in the two tests were described in detail in Section 3. The systems are summarized in the following subsections.

#### 4.3.1 Test 1

The codecs and their properties in Test 1 are found in Table 6.
Table 6. Codecs in Test 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>System</th>
<th>Audio b/w [kHz]¹</th>
<th>Subchannel bit rate [kbit/s]</th>
<th>Audio bit rate (incl. PAD) [kbit/s]</th>
<th>PAD bit rate [kbit/s]</th>
<th>Spectral Band Replication (SBR)</th>
<th>Parametric Stereo (PS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAB+ 96 (SBR)</td>
<td>DAB+</td>
<td>22.4</td>
<td>96</td>
<td>87.2</td>
<td>1.3</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 128</td>
<td>DAB+</td>
<td>15.3-16.2</td>
<td>128</td>
<td>115.8</td>
<td>2.53</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 160</td>
<td>DAB+</td>
<td>15.3-16.2</td>
<td>160</td>
<td>145.1</td>
<td>2.53</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 192</td>
<td>DAB+</td>
<td>17.0-19.5</td>
<td>192</td>
<td>174.5</td>
<td>2.53</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

¹ Measured bandwidth. If two values occur, then most audio reaches the lower value while transients can reach the higher value.

4.3.2 Test 2

The systems and their properties in Test 2 are found in Table 7.

Table 7. Systems in Test 2.

<table>
<thead>
<tr>
<th>Name</th>
<th>System</th>
<th>Audio b/w [kHz]²</th>
<th>Subchannel bit rate [kbit/s]</th>
<th>Audio bit rate (incl. PAD) [kbit/s]</th>
<th>PAD bit rate [kbit/s]</th>
<th>Spectral Band Replication (SBR)</th>
<th>Parametric Stereo (PS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAB+ 48 (SBR, PS)</td>
<td>DAB+</td>
<td>18.7</td>
<td>48</td>
<td>43.2</td>
<td>1.3</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>DAB+ 64 (SBR)</td>
<td>DAB+</td>
<td>18.7</td>
<td>64</td>
<td>57.9</td>
<td>1.3</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 96 (SBR)</td>
<td>DAB+</td>
<td>22.4</td>
<td>96</td>
<td>87.2</td>
<td>1.3</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 128</td>
<td>DAB+</td>
<td>15.3-16.2</td>
<td>128</td>
<td>115.8</td>
<td>2.7</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 160</td>
<td>DAB+</td>
<td>15.3-16.2</td>
<td>160</td>
<td>145.1</td>
<td>2.7</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DAB+ 192</td>
<td>DAB+</td>
<td>17.0-19.5</td>
<td>192</td>
<td>174.5</td>
<td>2.7</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>FM 1</td>
<td>FM stereo</td>
<td>17.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FM 2</td>
<td>FM stereo</td>
<td>15.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anchor</td>
<td>24.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hidden reference</td>
<td>24.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

² Measured bandwidth. If two values occur, then most audio reaches the lower value while transients can reach the higher value.

4.4 Equipment and listening environment

The listening equipment was selected to match the previous BPN 094 test (EBU, 2009) (ITU, 1990). Headphones were used to reduce the possible differences between room characteristics of the two test sites. The headphone amplifier Grace Design m903 contained a D to A converter using an accurate internal clock. Diffuse-field EQ was employed to achieve the correct frequency response in the headphones (Sennheiser HD-650). The filter characteristics were developed by the IRT for BPN 094 and were adapted for the HD-650 headphones specifically. The filter characteristics were employed in these tests by processing the audio files using convolution at a high resolution (ITU, 1990). This amplifier was connected to the STEP software⁹, version 1.08a, via a USB interface. Jitter performance was measured in the analogue domain with the Audio Precision System Two jitter signal. The values were found to be in the same order of magnitude or better than other high-performance converters.

⁹ www.audioresearchlabs.com
The noise levels of the listening rooms were well under NR 15 at Piteå and under NR 20 in Stockholm.

4.5 Training
Subjects invited to the listening test were sent a link to audio files that they were asked to listen to at home for training purposes before the tests. The audio files contained all audio that the subjects were to grade in the actual tests. Everything that was not the original audio was relabelled as random numbers so as not to give any clues of the systems they had passed through. Every subject received and was asked to use two randomized playlists for the audio files. If some subjects used playback equipment of varying quality or choose not to listen to all audio files, this procedure would distribute the training sessions randomly over all audio files.

The actual tests were preceded by training that consisted of grading a random subset of the audio that was to be graded later in the actual test. This way each subject learned the functions of the user interface, became familiar with the listening environment and equipment, and practiced grading some of the audio from the actual test. The training for Test 1 (BS.1116) consisted of a subset comprising five gradings, randomly selected for each subject. Each of the five excerpts appeared once, three of the four bit rates appeared once, and one bit rate appeared twice. The training for Test 2 (MUSHRA) consisted of a randomly selected subset of three excerpts processed in all versions. Each subset was randomly selected to distribute the training randomly over all subjects.

4.6 Procedure
Two test leaders worked together during the test. At the time of the tests, each subject received the same instructions and everyone was instructed by the same two test leaders. In almost all cases, both test leaders were present. In this way, the two test leaders complemented each other to ensure that all subjects received very similar oral instructions before the test. The test leaders used a checklist to make the instructions as consistent as possible. The test was totally blind to the listeners. The test leaders were careful not to influence the subjects in how they graded different artefacts and the type of systems or codecs tested were never mentioned.

Each subject first trained for Test 1 (BS.1116) before performing Test 1 and trained for Test 2 (MUSHRA) before performing Test 2. Normally, a break with coffee or tea and sandwiches was provided after the training for Test 2.

The subjects were allowed to adjust the listening levels at any time during the tests. However, the loudness alignment performed before the experiment seemed to remove possible loudness differences as only very few and small adjustments were made by the listeners.

4.6.1 Test 1
The following information and instructions were given to the listeners before the BS.1116 test:
1. A general background to the test.
2. An overview of the test methodology and user interface.
3. To take regular breaks.
4. Usage of the looping functionality was only allowed when a subject had listened to all audio from beginning to end.
5. Subjects were informed about the level equalisation applied. They were encouraged to set a comfortable listening level and were also allowed to adjust the listening level at any time during the test.
6. The audio quality change as a whole – i.e., “basic audio quality” – should be considered when deciding on a grade, which means that all audio quality differences from the reference should be considered to be audio quality degradations.
7. Only the degradations in audio quality should be considered when deciding on a grade and not how good the mix is or how the listener enjoys the content.
8. The listener should be careful not to pull the wrong slider when giving a grade.
9. It is not possible to go back and change a previously given grade.
10. Listeners should make a note of the listening levels they used.

4.6.2 Test 2
The following information and instructions were given to the listeners before the MUSHRA test:

1. The same instructions as for Test 1, but now the test methodology and user interface described the MUSHRA test.
2. The listener can compare all processed versions of the audio with the reference and also comparisons can be made between the different processed versions of the audio.
3. The audio in Test 2 is normalized to a lower level than for Test 1.

4.6.3 Interviews
Interviews were conducted (one interview per subject) after the listening tests. For the purpose of this paper, four of the interviews, two from each test site, were randomly selected for analysis. The methodology used had a phenomenological approach for identifying the subject’s experience (Merriam, 2009). In this study, this refers to the underlying structure of what degradations the subjects perceived and how the degradations were perceived. An implication of applying a qualitative method is that the results cannot necessarily be applied to the general public, but it gives an insight into understanding the subjects’ experience and the tested audio’s perceived qualities. Thus, it can complement the statistical results.

An interview guide was used to guide the interviewer, and open-ended questions were asked to encourage descriptive statements from the subjects. The questions included the following topics: perceived degradations, what gave low and high scores, the listening training at home, differences between the two tests
(BS.1116 and MUSHRA), perception of the scales used in the tests, looping habits, and perceived sound level between the tests.

4.7 Data analysis
The data analysis mainly follows the recommendations of the ITU for such tests (ITU, 1997) (ITU, 2003). Additional analysis is also performed to complete the research objectives. The analysis methods are briefly described in this section. Section 5 presents the results and analysis.

4.7.1 Test 1
In Test 1, analysis of variance (ANOVA) was used. The factors in the model were as follows: Codec = Coding process under test (4 levels, Table 6); Excerpt = Sound excerpt (5 levels, Table 4); and TestSite = Site where test was performed (two levels, Piteå/Stockholm). The ANOVA model with the dependent variable Subjective Difference Grade (SDG) was determined using the following equation:

\[ \text{SDG} = \text{Mean} + \text{Codec} + \text{Excerpt} + \text{TestSite} + \text{Codec*Excerpt} + \text{Codec*TestSite} + \text{Excerpt*TestSite} + \text{Codec*Excerpt*TestSite} + \text{Error}. \]

Post hoc tests were also made to investigate differences within the factors Codec and Excerpt. Additionally, the acquired data were used to find a model for perceived audio quality as a function of bit rate, assuming a linear relation between those. Finally, data from Test 1 were mapped against corresponding data from a previous test within EBU, project D/DABA (EBU, 2009), to observe the relation between the scales of the BS.1116 and MUSHRA evaluation methods. This is reported in a separate section (5.5).

4.7.2 Test 2
In Test 2, mean values of the different conditions together with their associated 95% confidence intervals were calculated. These were investigated for significant differences and compliance to the transparency criteria in Section 2. In addition, a number of planned comparisons between FM and DAB+ were made by means of pair-wise t tests, where the effect of multiple comparisons were controlled by applying the false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995).

4.7.3 Interviews
The analysis used meaning condensation as the primary method, where large segments of descriptive transcripts are condensed into short units of meaning, or “the essence” (Kvale & Brinkmann, 2009) (Tesch, 1990). The units of meaning are then labelled into common themes for all four interviews and then presented as descriptive texts.

5 Results and analysis

5.1 Test 1
For each trial, the Subjective Difference Grade (SDG) was calculated by subtracting the grade assigned to reference from the grade assigned to the object under test.
5.1.1 Post-screening
Post-screening identifies and rejects subjects that show an inability to discriminate the reference (unprocessed) signals. The procedure follows the recommendations in BS.1116. First, any item that received low average grade across all subjects, i.e., having a mean SDG below ~2.0, was temporarily removed during the following stage. Second, for every subject, the remaining data thus obtained was subjected to a one-sided t test ($\alpha=0.05$) to assess the likelihood that the mean of the distribution for each subject is zero or greater. Subjects failing to produce a mean that is significantly less than zero at $\alpha=0.05$ were rejected from the subsequent analysis. As a result of this process, nine subjects were rejected and the associated data were removed from the data set. Hence, data from 21 subjects remained for analysis. After this adjustment, the data set contained 420 data points (21 subjects * 4 codecs * 5 items).

5.1.2 ANOVA
The ANOVA residuals were tested for normal distribution using the Kolmogorov-Smirnov test. As $K=0.038$ ($p=0.166$), normal distribution of residuals could not be rejected and is therefore assumed. In addition, for each combination of codec and excerpt, the data were checked for normal distribution by means of the Shapiro-Wilk test. Out of the 20 combinations, the null hypothesis was rejected for five cases only ($\alpha=0.05$). This implies that the vast majority of the data come from a normally distributed population. Levene’s test of equality of error variances yielded $F(39,380)=1.227$ ($p=0.172$). Hence, equal error variances could not be rejected and were therefore assumed. The experimental design included a randomization of both trial order and assignment of stimuli to the graphical user interface buttons within each trial. Thus independency between data points was ascertained. In summary, the assumptions underlying ANOVA were not violated. The ANOVA is summarised in Table 8.
Table 8. ANOVA Table.

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>421.883</td>
<td>39</td>
<td>10.818</td>
<td>15.619</td>
<td>.000</td>
<td>.616</td>
</tr>
<tr>
<td>Intercept</td>
<td>991.737</td>
<td>1</td>
<td>991.737</td>
<td>1431.943</td>
<td>.000</td>
<td>.790</td>
</tr>
<tr>
<td>Codec</td>
<td>117.840</td>
<td>3</td>
<td>39.280</td>
<td>56.715</td>
<td>.000</td>
<td>.309</td>
</tr>
<tr>
<td>Excerpt</td>
<td>253.043</td>
<td>4</td>
<td>63.261</td>
<td>91.340</td>
<td>.000</td>
<td>.490</td>
</tr>
<tr>
<td>TestSite</td>
<td>7.724</td>
<td>1</td>
<td>7.724</td>
<td>11.152</td>
<td>.001</td>
<td>.029</td>
</tr>
<tr>
<td>Codec * Excerpt</td>
<td>12.097</td>
<td>12</td>
<td>1.008</td>
<td>1.456</td>
<td>.139</td>
<td>.044</td>
</tr>
<tr>
<td>Codec * TestSite</td>
<td>2.371</td>
<td>3</td>
<td>.790</td>
<td>1.141</td>
<td>.332</td>
<td>.009</td>
</tr>
<tr>
<td>Excerpt * TestSite</td>
<td>3.862</td>
<td>4</td>
<td>.965</td>
<td>1.394</td>
<td>.235</td>
<td>.014</td>
</tr>
<tr>
<td>Codec * Excerpt * TestSite</td>
<td>8.175</td>
<td>12</td>
<td>.681</td>
<td>.984</td>
<td>.464</td>
<td>.030</td>
</tr>
<tr>
<td>Error</td>
<td>263.181</td>
<td>380</td>
<td>.693</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1723.150</td>
<td>420</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>685.064</td>
<td>419</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. R Squared = .616 (Adjusted R Squared = .576)

The ANOVA showed that all three main factors were significant. The largest effect size was found for Codec (η²=0.31) and Excerpt (η²=0.49). For the remaining factors and combinations, the effects were negligible (η²<0.05), so the factors Codec and Excerpt were further investigated and subjected to post-hoc tests.

5.1.3 Multiple comparisons of codecs
The SDG mean values and their associated confidence intervals for different codecs across all excerpts are shown in Figure 7. A Tukey HSD post-hoc test (α=0.05) was performed to find the significant differences between codecs (Table 9). The results showed that there is a significant difference between the mean SDG for every pair-wise combination of the codecs. The mean SDG rises significantly for each increase in bit rate.
Figure 7. Mean values and 95% confidence intervals for codecs across excerpts.

Table 9. Multiple comparisons of codecs (Tukey HSD); mean SDGs and resulting homogenous codec subsets (CS).

<table>
<thead>
<tr>
<th>Codec</th>
<th>N</th>
<th>CS1</th>
<th>CS2</th>
<th>CS3</th>
<th>CS4</th>
</tr>
</thead>
<tbody>
<tr>
<td>96 kbit/s (AAC + SBR)</td>
<td>105</td>
<td>-2.303</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>128 kbit/s (AAC)</td>
<td>105</td>
<td></td>
<td>-1.820</td>
<td></td>
<td></td>
</tr>
<tr>
<td>160 kbit/s (AAC)</td>
<td>105</td>
<td></td>
<td></td>
<td>-1.313</td>
<td></td>
</tr>
<tr>
<td>192 kbit/s (AAC)</td>
<td>105</td>
<td></td>
<td></td>
<td></td>
<td>-0.852</td>
</tr>
<tr>
<td>Sig.</td>
<td></td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

The error term is Mean Square (Error) = 0.693.

5.1.4 Multiple comparison of excerpts
The SDG mean values and their associated confidence intervals for different codecs and excerpts are shown in Figure 8. A Tukey HSD post-hoc test ($\alpha=0.05$) was performed to find the significant differences between excerpts across codecs (Table 10). The results showed that there were three excerpt subsets (ES) where the excerpts within each subset were not significantly different from one another, but differentiated significantly from the excerpts of the other subsets. For each subset, the content was as follows: ES1 – Speech(pan) and Speech(pan)+5xL2; ES2 – House; and ES3 – Applause w announcer and Classical.
Figure 8. Mean SDG and 95 % confidence intervals for codecs and excerpts.

Table 10. Multiple comparisons of excerpts (Tukey HSD); mean SDGs and resulting homogenous excerpt subsets (ES).

<table>
<thead>
<tr>
<th>Excerpt</th>
<th>N</th>
<th>ES1</th>
<th>ES2</th>
<th>ES3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech (pan)</td>
<td>84</td>
<td>-2.554</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech (pan) + 5xL2</td>
<td>84</td>
<td>-2.496</td>
<td></td>
<td></td>
</tr>
<tr>
<td>House</td>
<td>84</td>
<td></td>
<td>-1.190</td>
<td></td>
</tr>
<tr>
<td>Applause w announcer</td>
<td>84</td>
<td></td>
<td></td>
<td>-0.830</td>
</tr>
<tr>
<td>Classical</td>
<td>84</td>
<td></td>
<td></td>
<td>-0.790</td>
</tr>
<tr>
<td>Sig.</td>
<td>84</td>
<td>0.992</td>
<td>1.000</td>
<td>0.998</td>
</tr>
</tbody>
</table>

The error term is Mean Square (Error) = 0.693.

5.2 Test 2

5.2.1 Post-screening
The post-screening was performed according to step 1–4 below and its objective was to reject subjects that were unable to discriminate between stimuli in terms of reference signals and anchor signals. The score range used by each subject was also investigated.
1. **Ability to identify reference signals.** The probability of correctly identifying the hidden reference signal by chance in one trial including ten signals is 0.1. The total number of trials is ten. The minimum required number of correct identifications for all trials, \( n_{ID} \), at \( \alpha = 0.01 \) is calculated by means of the cumulative distribution function of the binomial distribution satisfying \( b(x; 10, 0.1) \geq 0.99 \), where \( n_{ID} > x \). Consequently, a subject must show \( n_{ID} \geq 5 \) to fulfil this condition. When applying this condition, seven subjects were rejected from the subsequent analyses.

2. **Ability to score reference signals.** To ascertain that subjects passing the test above do not underscore the reference signals, reference signals where Score<90 were counted. Subjects showing more than three such scores were rejected. Five subjects failed this test; four of them were already rejected in the previous step. Hence, one additional subject was rejected as a result of this.

3. **Ability to identify and score anchor signals.** To ensure the assessment of anchor signals, two measures were used for each subject: a) the mean value of Anchor scores and b) the number, \( n_{EAS} \), of scores exceeding a fixed score value. The value used in both measures was Score=60. As a score above this value would represent “good” or “excellent”, it is reasonable not to expect such a quality for a majority of Anchor signals. In b), \( n_{EAS} = 5 \) was chosen. When applying any of these conditions, one subject failed, but was already rejected in step 1 above.

4. **Score range used by subjects.** To accommodate for the recommendation on “less critical” and/or “too critical” subjects, the remaining subjects’ scores were analysed for each combination of codec and excerpt, yielding 100 combinations. The interquartile range, IQR, was calculated for the scores of each combination. A score outside the range 3IQR from the median was considered an extreme value. The number of extreme values per subject was checked and in no case did any subject show a number exceeding 4 out of the 100 judgements made. This was considered a valid performance. Hence, no subjects were rejected on this criterion only.

In total, eight subjects were rejected as a result of post-screening. The data set now contained 2200 data points (22 subjects * 10 systems * 10 items).

### 5.2.2 Multiple comparisons of systems

In this section, the performance of the different systems across all excerpts except the Electronic (att) is shown (Figure 9). The Electronic (att) data was excluded from the figure because this excerpt was inserted into the systems at an attenuated level and therefore did not reflect normal broadcasting operating conditions. The results for this excerpt (Section 5.2.3.4) confirm that it was perceived differently from the other excerpts by exhibiting low scores on all systems due to a high noise level and thus would be unfair to include. In Figure 9, clear differences can be seen between the different conditions. In the following sections, the differences will be investigated further.
5.2.3 Multiple comparisons of system*excerpt
In this section the performance of the different systems for each excerpt is shown (Figure 10 through Figure 19) together with a brief analysis on possible explanations to system performance.
5.2.3.1 Applause with announcer

![Graph showing mean scores and 95% confidence intervals for different codecs.]

Figure 10: Mean values and 95% confidence intervals for Excerpt = Applause with announcer.

The original audio was slightly band-limited, which may be the cause of the similar scores at the higher bit rates. It also had impulses that contained large amounts of high frequencies that were clipped in the FM systems. The difference between FM 1 and FM 2 was probably caused by the difference in method of clipping. The traditional pre-emphasis clipping on the left/right audio in FM 2 was probably the cause of its low score. In FM 1, the pre-emphasis clipping was done on the multiplex signal.
5.2.3.2 Classical

![Diagram showing mean values and 95% confidence intervals for Excerpt = Classical.]

Figure 11. Mean values and 95% confidence intervals for Excerpt = Classical.

The excerpt contained brass instruments and castanets with large amounts of high frequencies and transients clipped in the FM systems. As with the previous excerpt, the difference between FM 1 and FM 2 was most likely attributable to the clipping method.
5.2.3.3 Electronic

Figure 12. Mean values and 95% confidence intervals for Excerpt = Electronic.

The original audio contained quite strong and short transients, which means that they contained a lot of high frequency components. These components were clipped by the pre-emphasis clippers in the FM systems, a method that probably resulted in the low scores. In contrast to the two previous excerpts, FM 2 received a higher score, which implies that the artefacts were perceived as less severe for this excerpt.
5.2.3.4 Electronic (attenuated)

![Graph showing mean scores and 95% confidence intervals for different conditions]

**Figure 13.** Mean values and 95% confidence intervals for Excerpt = Electronic (attenuated).

This was the same excerpt as Electronic but attenuated by additionally 22 dB to avoid clipping the FM systems. Unfortunately, the make-up gain applied to this condition resulted in a high noise level that caused low scores for all systems. The main noise source was the FM exciter.
5.2.3.5 **House**

![Graph showing mean scores and 95% confidence intervals for Excerpt = House.](image)

**Codec**

Figure 14. Mean values and 95% confidence intervals for Excerpt = House.

This excerpt included a bass synthesizer with a complex waveform that contained short spikes that probably caused the low scores at lower DAB+ bit rates.
5.2.3.6 *PopKent*

![Graph showing mean scores and 95% confidence intervals for different codecs.](image)

**Codec**

*Figure 15.: Mean values and 95% confidence intervals for Excerpt = PopKent.*

The excerpt included a panned hi-hat that may have contributed to the low scores at the lower DAB+ bit rates. Except for the hi-hat, the excerpt did not include too much high frequency energy, a condition that proved to be advantageous for FM as well as for AAC.*
5.2.3.7 PopRox

![Figure 16. Mean values and 95 % confidence intervals for Excerpt = PopRox.](image)

This excerpt had a large amount of high frequency content. The systems employing Spectral Band Replication, SBR, yielded a higher audio bandwidth, which was most noticeable at 96 kbit/s.
5.2.3.8 *SpeechL (no pan)*

![Figure 17. Mean values and 95% confidence intervals for Excerpt = SpeechL (no pan).](image)

The female speech excerpt was recorded in mono at close distance. Because the voice in this excerpt was centred (equal signals in left and right channel), there was no difference signal (=L-R) that had to be coded. Hence, the AAC could allocate all bits to the sum signal. This was probably the reason for the high DAB+ scores, except for 48 kbit/s.
5.2.3.9 **SpeechL (pan)**

![Figure 18](image-url)

Figure 18. Mean values and 95% confidence intervals for Excerpt = SpeechL (pan).

The original audio contained only a single female speaker panned to one side: the interchannel level difference was 6 dB. In this case, using parametric stereo at 48 kbit/s had an advantage. At higher bit rates, parametric stereo was not used, which caused a blur in the stereo image. The traditional pre-emphasis clipping on the left/right audio in FM 2 was probably the cause of the low score. In FM 1, the pre-emphasis clipping was done on the multiplex signal.
5.2.3.10 World

Figure 19. Mean values and 95% confidence intervals for Excerpt = World.

For the World excerpt, the scores followed what may be expected from an increased bit rate.

5.2.3.11 Summary of multiple comparisons

The data summarised in Figure 10 through Figure 19 produced the following observations:

- In most cases, at higher bit rates no significant differences were observed, i.e., the confidence intervals overlap. The MUSHRA method may be less sensitive to the smaller differences at high bit rates.
- In several cases, there was confusion between reference and coded signals.
- Certain signals were more critical, i.e., panned speech and transient sounds also occurring in modern music.
- Monotonically increasing scores for increasing bit rate were observed with a few exceptions. This was probably partly caused by the general increase in audio bandwidth and decrease in quantization distortion that follows with higher bit rate.
- FM 1 outperformed DAB+ at low bit rates.
- FM 2 was more sensitive to certain excerpts, probably due to its different clipping algorithm in combination with its narrower bandwidth.
5.3 Compliance to audio quality criteria
To reiterate, four criteria of audio quality and transparency were discussed in Section 2. If bit transparency is omitted, as it was not aimed for in any of the tested systems, three criteria remain:

1. perceptual transparency, which means that no statistically significant difference between the reference and a tested system should be found;
2. broadcast quality, which means that the mean across all items, and preferably also the mean of each item, should fulfil SDG > -1 in BS.1116 or Score > 80 in MUSHRA; and
3. FM quality, which means that the tested system should show equal quality when compared with a specific FM system.

The compliance of the tested systems to the criteria was investigated and these investigations are discussed below.

5.3.1 Perceptual transparency and broadcast quality
In Test 1, none of the systems was found to be perceptually transparent. For the bit rates 96, 128, and 160 kbit/s, none of the tested items fulfilled this criterion. At 192 kbit/s, the items “Applause w announcer” and the two speech items did not reach perceptual transparency. The broadcast quality criterion was fulfilled in Test 1 only at 192 kbit/s, but since panned speech at this bit rate did not reach a SDG > -1, this system should be used with caution. For Test 2, a similar analysis was made (Table 11).

Table 11. Compliance to quality criterion 1 and 2 for the different systems in Test 2.

<table>
<thead>
<tr>
<th>Excerpt</th>
<th>DAB+</th>
<th>FM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>48</td>
<td>64</td>
</tr>
<tr>
<td>Applause w announcer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electronic (att)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>House</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PopKent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PopRox</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SpeechL (no pan)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SpeechL (pan)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>World</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean value for systems*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Electronic (att) excluded
As is shown by the horizontal lines in the table, none of the tested systems were fully perceptually transparent. At 192 kbit/s, all tested items except Electronic (att) and panned speech, however, fulfilled this criterion. As mentioned previously, the Electronic (att) excerpt does not reflect normal operating conditions and should be disregarded, whereas panned speech is commonly used in broadcasting and therefore should be considered an important excerpt.

The vertical lines in Table 11 show that both 192 kbit/s and FM 1 in Test 2 fulfilled the broadcast quality criterion. The average score across items for these systems was above 80, but since some individual items score less than 80, these systems should be used with care according to the discussion on broadcast quality in Section 2.

5.3.2 FM quality
To test specifically the difference between the systems DAB+ and FM, paired two-tailed t tests (df=21) were performed on the mean difference $d$ between the scores for DAB+ and FM1 for each of the bit rates of DAB+ systems and excerpts except Electronic (att). $d$=Score(DAB+, [bit rate, excerpt])-Score(FM 1, [excerpt]). This was repeated for FM 2 replacing FM 1. The resulting p values from all t tests were subjected to the FDR procedure for declaring significant differences based on $\alpha = 0.05$ (Benjamini & Hochberg, 1995).

For FM 1 (Table 12), the results showed that in only two out of the nine cases at 192 kbit/s the DAB+ system was perceived as significantly better, whereas at lower bit rates either no difference or worse performance by the DAB+ was noted. For FM 2 (Table 13), the results were more dispersed: at higher bit rates, DAB+ surpassed FM 2; at lower bit rates, DAB+ fell below; and in a number of cases, no significant differences were found. There were instances where DAB+ was superior to FM 2 down to 64 kbit/s but also inferior up to 128 kbit/s.

Table 12. t tests of differences between DAB+ and FM 1 for excerpts except Electronic (att). Significant differences are denoted by “+” for a higher score (“-“ for a lower) of DAB+ systems.

<table>
<thead>
<tr>
<th>Excerpt</th>
<th>DAB+ systems at different bit rates compared with FM 1 (kbit/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>48</td>
</tr>
<tr>
<td>Applause w announcer</td>
<td>-</td>
</tr>
<tr>
<td>Classical</td>
<td>-</td>
</tr>
<tr>
<td>Electronic</td>
<td>-</td>
</tr>
<tr>
<td>House</td>
<td>-</td>
</tr>
<tr>
<td>PopKent</td>
<td>-</td>
</tr>
<tr>
<td>PopRox</td>
<td>-</td>
</tr>
<tr>
<td>SpeechL(no pan)</td>
<td>-</td>
</tr>
<tr>
<td>SpeechL(pan)</td>
<td>-</td>
</tr>
<tr>
<td>World</td>
<td></td>
</tr>
</tbody>
</table>
Table 13. t tests of differences between DAB+ and FM 2 for excerpts except Electronic (att). Significant differences are denoted by “+” for a higher score (“-” for a lower) of DAB+ systems.

<table>
<thead>
<tr>
<th>Excerpt</th>
<th>48</th>
<th>64</th>
<th>96</th>
<th>128</th>
<th>160</th>
<th>192</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applause w announcer</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Classical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electronic</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>House</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>PopKent</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PopRox</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>SpeechL(no pan)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SpeechL(pan)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>World</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.3.3 Bit rates required for transparency

To predict the bit rate necessary to attain SDG=0 (i.e., imperceptible difference from the reference), a model for SDG as a function of bit rate, \( R_b \), in the form of a linear curve, \( SDG = b R_b + c \), was fitted onto the data in Test 1. As the excerpts formed the three subsets (see Section 5.1.4) ES1 (SpeechL excerpts), ES2 (House excerpt) and ES3 (Applause and Classical excerpts), these were treated individually. The resulting parameters (\( b \) and \( c \)), the goodness-of-fit (\( R^2 \)), as well as the required bit rate for fulfilling SDG=0 was calculated for each of the subsets (Table 14).

Table 14. Parameters of the linear model \( SDG = f(R_b) \) and required bit rate for SDG=0.

<table>
<thead>
<tr>
<th>Excerpt subset</th>
<th>( b )</th>
<th>( c )</th>
<th>( R^2 )</th>
<th>( R_b ) for SDG=0 [kbit/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES1</td>
<td>0.018</td>
<td>-5.13</td>
<td>0.43</td>
<td>285</td>
</tr>
<tr>
<td>ES2</td>
<td>0.019</td>
<td>-3.93</td>
<td>0.35</td>
<td>207</td>
</tr>
<tr>
<td>ES3</td>
<td>0.010</td>
<td>-2.30</td>
<td>0.16</td>
<td>230</td>
</tr>
</tbody>
</table>

The results showed that the linear model pointed towards necessary bit rates above 200 kbit/s for any of the subsets to reach SDG = 0. The highest bit rate required was found for subset ES1, which contained the most critical items, where SDG = 0 was reached at \( R_b = 285 \) kbit/s. This subset also yielded the best fit (\( R^2=0.43 \)) of the model.

5.4 Interviews

The resulting themes from the interviews included information about the perceived degradations, which of them were prominent, whether they were easy/hard to perceive, and what degradations gave low/high scores during the tests. The analysis produced the following descriptive text.

The perceived degradations were changes in the frequency response, comb filter effects, low bit rate artefacts, changes in the stereo image, longer attack time of transients, more noise, and the degradation of the “feeling” of the sound. The most prominent degradations were changes in the stereo image and changes in
the frequency response. Degradations caused by noise were easy to perceive, e.g. general noise levels, static noise in silent parts of the stimuli. Also changes in the stereo image were prominent. Degradations were perceived when the voice’s stereo panorama position was off-centre, as well as when panned transients occurred.

Degradations were hard to perceive in complex stimuli that did not contain any easily discernable reference points. Some subjects focused on the general audio quality, including listening to the attack of the transients and stereo image to discern the quality of that stimulus. When distributing high scores, the subjects listened for the presence of a sound; one listener associated this with listening to a mix (as if an audio engineer). A sort of weighing strategy was also employed where the subjects listened for the total quality of the degradations to discern whether they were acceptable or whether the stimuli contained other advantages.

Several degradations were perceived as acceptable, e.g., a low amount of artefacts induced by the coding algorithm and loss of attack in the transients. One listener even stated that losses of high frequencies are more acceptable compared to coding artefacts, quantizing noise, and low bit depth.

Low scores were given when the following conditions were met:
- a large difference between the stimulus and the reference;
- degradations were audible in combination with how much it affected the result; e.g., speech was not scored the lowest because the information in the speech still came through to the listener;
- the bandwidth was extremely limited and when the bandwidth changed with the signal; and
- loss of high frequencies, high frequency tones, and loss of attack of the transients were perceived.

5.5 Relation between BS.1116 and MUSHRA scales
To find the relation between the scales of the BS.1116 and MUSHRA evaluation methods, data for 12 conditions from Test 1 were mapped against the same conditions from a previous EBU test (EBU, 2009). The variables were Grade, \( G \) (from BS.1116), and Score, \( S \) (from MUSHRA). Curve fits for a linear relation as well as other functions were performed. The linear curve fit \((R^2=0.93)\) resulted in \( G = 0.049S + 0.134 \). The best non-linear fit \((R^2=0.94)\) was the power function \( G = 0.055S^{0.983} \).

The models indicated a linear or close to linear relationship with no or small offset from the origin (Figure 20). It should be noted that in BS.1116, Grade is not defined below 1.0, which according to these models would mean that qualities at the lowest parts of the MUSHRA scale, below Score = 20, corresponding to “bad quality”, are not possible to represent on the BS.1116 scale.
Figure 20. Relation between BS.1116 Grade from Test 1 and MUSHRA Score from the EBU test BPN 094 (EBU, 2009) for the same items.

6 Discussion

6.1 Audio quality and transparency

Considering the results of the experiments in relation to the transparency criteria presented in Section 2, several observations can be made.

The preferred subchannel bit rate for the items in Test 2 transmitted over DAB+ should be at least 192 kbit/s for “broadcast quality” (as single items below Score=80 were allowed). For “perceptual transparency”, even higher bit rates would be required due to the lack of compliance of the SpeechL(pan) item. Such bit rates would pose a problem as DAB+ currently only allows for a maximum of 192 kbit/s.

If the results of Test 1 are considered and the definition for broadcast quality is applied, 192 kbit/s would be required, which supports the findings above. Still it has to be noted that the panned speech excerpts did not reach a sufficient level of quality. If the perceptual transparency criterion is applied and the linear model in 5.3 is used, this points towards a necessary bit rate above 200 kbit/s, and for the most critical items (panned speech), a bit rate close to 300 kbit/s (Table 14 indicates at least 285 kbit/s) would be needed. It has to be pointed out that the extrapolation by means of a linear function may contain errors that can affect the
accuracy of the predicted necessary bitrate. However, there is a need for bit rates to be well above 192 kbit/s, especially for critical items.

In Test 2, if the occurrence of a significant negative difference between DAB+ and FM for any item was to be applied as a criterion for non-transparency of DAB+ in relation to FM, the following would apply: a DAB+ bit rate of 192 kbit/s would give a result that is comparable to or better than a modern FM system (FM 1), whereas a bit rate of 160 kbit/s is likely to perform comparable to or better than the average types of FM transmitters used by Swedish Radio (FM 2). Lower bit rates would give rise to significant degradations of the audio quality (Table 12 and Table 13).

Certain types of sounds appear quite commonly in regular programming, e.g., applause and panned speech in radio. The quality of such sounds will have a significant weight on the overall perceived quality of a broadcasting channel and will therefore be essential to include in quality tests. The occurrence of such signals in a test is also important to consider when averaging scores across items as results from one quite critical item may be masked by less critical ones. Any representation of data from a listening test that just shows the average performance across items may risk missing some of the quality deficiencies of the systems. One example is depicted in Figure 9 where DAB+ at 128 kbit/s appears to be better than FM 2 on average. However, the analysis of individual items (Table 13) shows that two items have significantly inferior quality. Although the findings in the current study were based on a very limited number of excerpts, they show that sounds that are critical exist as they revealed weaknesses of the systems under test.

A well-known phenomenon is that a codec may have performed well for a vast number of sound excerpts over time, but when exposed to a previously never encountered excerpt, it may produce clearly audible artefacts at the selected bit rate. To reduce such risks when a system’s bit rate is to be established and to accommodate for future possible critical items, one solution may be to apply a safety margin in the form of a bit rate that is higher than the one the current listening test indicates as being necessary.

Although the experimental design was aimed to emulate realistic broadcasting conditions, it should be noted that the experiments were performed under more favourable conditions than are normally found in real-life broadcasting. One example is that mobile reception over a long distance from the transmitters may degrade the audio in various ways, both for FM and DAB+. Another example is the common use of cascaded codecs, especially if different bit rates are used at different stages of the cascade. Also, ancillary data may gradually use more bandwidth and thus increase over time, which may in reality lead to a reduction of the available audio bit rate.

It should also be noted that the DAB+ and FM processes were limited to a resolution of 16 bits at the time of the experiments, whereas systems with higher resolution are now available. Another observation about listening tests in
general is that some of the reference signals still in use may have been recorded through equipment that now has been surpassed in terms of quality.

As noted previously in this paper, in comparison with a number of other media distribution formats, the bit rates currently used for DAB+ are generally lower. An obvious conclusion is that DAB+ listeners could perceive the system as inferior to other contemporary audio applications.

Altogether the experimental results imply that even higher bit rates may be needed to reach the desired level of quality and transparency in future broadcasting systems.

6.2 Miscellaneous observations
From the interviews, a number of features important for the subjects’ assessment of quality were observed. In addition to statements on weighing coding artefacts against reduced audio bandwidth, several details were reported. One particularly interesting detail was related to the appearance of artefacts at stereo panorama positions segregated from the voice’s position. Other aspects outside the scope of this paper were also observed, e.g., comparison and evaluation strategies used by the listeners as well as how the scales used were perceived. Observations were also made on how effective the training process at home was seen from the subjects’ point of view as well as on the strategies the subjects used to rank the excerpts. More details on these findings will be presented in the future.

The excerpt Electronic (att) was attenuated so as not to cause clipping in the pre-emphasis circuit. The idea was to compare the excerpt with its un-attenuated counterpart (Electronic). However, due to the low signal level and the following make-up gain, the noise level became so loud that the noise itself caught the listeners’ attention, which led to significantly lower scores for the attenuated excerpt. This did not correspond to normal broadcasting operating conditions. Additionally, some of the possible artefacts may have been masked by the noise. Consequently, the intended comparison was not possible and had to be abandoned.

In Test 2, the distribution of scores was different between excerpts. In some cases, large quality differences were found; for other cases, the opposite was observed. When the differences are small due to a perceived high quality of the items, the resolution of the MUSHRA method may be insufficient. On the other hand, tests where codecs are compared indirectly by means of a reference, such as the BS.1116, would result in problems where items have large perceptual differences (Soulodre & Lavoie, 1999).

“FM quality” has been used in a wide range of meanings by the audio community. In this experiment, measures were taken to ensure a high-quality performance within the possible limits. This was accomplished by employment of the recommendations ITU-R BS.412-9 regarding the average multiplex power
deviation as well as the peak deviation and EBU R-128 regarding audio levels (ITU, 1998) (EBU, 2010).

As discussed in 3.2.4, band limited anchor signals are normally used in MUSHRA tests. Typically, the anchor is scored quite low while the other excerpts receive scores that are relatively high in comparison to the anchor due to the anchor’s large perceptual difference from the wide-band items. In this experiment, the use of a wide-band frequency spectrum anchor removed some of these effects for several items.

The relation between the BS.11 16 and the MUSHRA evaluation scales exhibits a linear or close to linear behaviour for the investigated conditions. Although such a relation is far from straightforward, it gives some support to the sometimes discussed similarity between Grade=4 (or SDG=-1) and Score=80. There are issues related to scale compression, ceiling effects, different biases, etc. that have to be investigated further before such similarities are adopted.

6.3 Future research
The quality issues in broadcasting are numerous and further research is needed in several areas. Some of them are summarized in this section.

Analogue FM is likely to be used for many more years in several countries. Further developments in loudness alignment and control of the FM deviation together with new transmitter equipment are expected to improve the FM quality performance even further (EBU, 2010) (ITU, 1998). This will reinforce FM as a de-facto anchor for audio broadcasting quality that subsequent systems will be compared with, which in its turn puts even more quality pressure on these systems. This fact in combination with the introduction of improved receivers both for FM and DAB+ as well as hybrid receivers for Web radio, DAB+, and FM will raise the question about how these systems relate to one another in terms of quality.

Antenna systems at the transmitter site and also receiver antennas need to be evaluated further. In United Kingdom and Italy, new national standards for certification of DAB and DAB+ receivers are in the process of being published. These new methods of measuring RF performance and functionality of digital radio receivers, however, do not take into account the perceived audio quality criteria as presented in this paper.

The current study is made with static reception. A long distance to the transmitter may degrade the audio in various ways. For FM, this results in multipath distortion, noise, etc.; for DAB, dropouts and other artefacts may be evident. More efficient error correction methods and ways to increase field strength, e.g., by means of transmitter power, may be needed, especially for mobile reception.

Signal loss measurements under realistic reception conditions would give an alternative interesting evaluation of the overall performance of a digital radio
system. What would be the listeners’ reactions to these types of artefacts?

More tests could clarify the impact of low bit rate systems for contribution and production in cascade with low bit rates used for distribution of digital radio. Test 2 showed that some of the effects of band limited anchor signals were suppressed when using a wideband anchor as defined in Section 3.2.4. This is a promising observation and comparison of traditionally used anchor signals with new types of anchor signals should be investigated further.

In summary, the development and refinement of systems for audio broadcasting calls for thorough testing and comparison of the systems in relation to other audio systems available. In such tests, as the findings in this paper show, it will be important to define criteria for acceptable quality and/or transparency. Hence, both test design and criteria for quality decisions are essential components of future research.

7 Conclusions
To conclude, the systems under test could not be considered as being fully transparent nor could they outperform a realistic optimal FM system on all accounts. The implications of the findings are listed below:

- The subchannel bit rate for DAB+ should not be less than 192 kbit/s for a stereo signal.
- A DAB+ subchannel bit rate of 192 kbit/s would be comparable to or better than the modern FM system.
- A DAB+ subchannel bit rate of 160 kbit/s would be comparable to or better than the average types of FM transmitters used by Swedish Radio.
- Bit rates below these could significantly degrade the quality of certain programme material.
- To accommodate for more critical but still typical items, unless encoding improves, a bit rate close to 300 kbit/s may be necessary for perceptual transparency to be realised.
- When making decisions about broadcasting systems, it will be important to have well-defined criteria for acceptable quality and whether perceptual transparency should be required.

Clearly, the bit rates and encoders in this study would have problems keeping up with the quality of other high-performance audio applications available to the end user. The available bit rates are subject to reduction due to both the number of competing channels that should be fitted as well as the ancillary data included in the bit stream. If broadcasting services advocate and market audio quality as their hallmark, interested parties need to make well-founded decisions on the audio coding infrastructure with respect to contribution, distribution, and emission.

8 Acknowledgements
The authors wish to thank the participating subjects in the listening test for their time and effort. Mr. Gerhard Spikofski (IRT) and Mr. Mathias Coinchin (EBU) are
thanked for providing data and support from the EBU D/DABA BPN 094 test. The
authors are grateful to Mr. Leif Claesson (Claesson-Edwards, Omnia Audio and
Linear Acoustic) and Teracom for providing FM equipment. Finally, Dr. Francis
Rumsey and Mr. Andrew Mason are acknowledged for their valuable comments.

9 References

powerful approach to multiple testing. Journal of Royal Statistical Society. Series B
(Methodological), 57 (1), 289-300.


EBU. (2009). Document BPN 094: Subjective Assessment and Objective measurements of

EBU. (2010). EBU Recommendation R 128: Loudness normalisation and permitted


ETSI. (2010). Digital Audio Broadcasting (DAB); Transport of Advanced Audio Coding


ITU. (1997). Recommendation ITU-R BS.1116-1: Methods for the subjective assessment of


sound broadcasting at VHF. Geneva: International Telecommunication Union.


Kvale, S., & Brinkmann, S. (2009). InterViews: Learning the Craft of Qualitative Research


Paper 3:
A Descriptive Method for Unravelling the Composition of the
Black Box of Affective Judgment in Sound Quality Evaluation

Submitted to: Journal of the Audio Engineering Society
A Descriptive Method for Unravelling the Composition of the Black Box of Affective Judgment in Sound Quality Evaluation

Dan Nyberg¹, Jan Berg¹

¹Luleå University of Technology

Abstract
This paper presents a descriptive (qualitative) method for unravelling a detailed composition of affective judgment of audio quality. Interviews are used to collect data and condensation was used as the analysis method. This descriptive method was used in a listening test and the results from it are the basis of the discussion about the method applicability. The descriptive method is shown to collect detailed information about the subjects’ affective judgment including the subjects’ moods, expectations, attitudes, listening habits, and strategies when preference rankings as well as the associations the subjects have to perceived differences.

1. Introduction
Typically, perceptual audio quality is analysed collecting data from three domains: the physical, the sensory, and the affective [2]. Several studies have been made on each domain separately as well as on the relation between the different domains. Affective measurements/judgments of audio quality, the focus of this study, includes the subjective preference of listeners. When affective judgments are conducted, the subjects (listeners) are assumed to combine context, expectations, previous experience, traditions, and any other attribute (e.g., spatial impression) in order to form an impression of the audio quality. The subject grades this judgment/measurement on a hedonic scale [1,2]. The data are then statistically analysed. Because any other deeper understanding of what goes into this judgment – e.g., how the subject feels, perceives, and understands the stimuli – are limited to the affective attribute under investigation, affective judgment can be seen as a “black box”.

Since it is assumed that there are several aspects that go into affective judgment, identifying what composes affective judgment needs further investigation. This shifts the focus of research from a representative population to the individual subject’s experience. Understanding what affective judgment is composed of can provide information on how naïve listeners judge sound quality. Bech and Zacharov describe this relationship as follows:

The subject (listeners) is the central measuring device that supplies the basic data to be analysed, and the more the knowledge about the “instrument”, the easier it is to understand the data it provides. ([2], P.106)
This study uses a purely descriptive\(^1\) (qualitative) method (phenomenological analysis of interviews) as a first step to unravel “the black box” of affective judgment. A listening test is employed where semi-structured interviews are conducted before and during the test as naïve subjects make affective judgments in the form of preference rankings for perceptually encoded and PCM coded audio. Interviews are chosen as the data collecting method since they let the subjects verbally describe their perception freely. The interviews also help subjects reflect on how their affective judgment works.

Rather than focusing on the results of the preference ranking, this study will focus on the following three issues:

- A qualitative analysis of the data;
- The type of information collected with this methodology; and
- The composition of the “black box”.

This study begins by providing background on material on affective judgment in perceptual audio research. Next, the study presents the framework of the descriptive research method. Then, the study presents the experimental setup, including an analysis of the methodology and results. Finally, the study discusses the methodology, the outcome, and the implications of the work.

2. Affective judgments
This section presents information about how affective judgments are conducted, focusing on the choice of subjects and what biases exist in affective judgment.

2.1 Listeners
When conducting affective measurement tests, un-trained/naïve subjects (listeners) are often used [2]. One reason for this is because they take less weight on spatial properties than trained subjects [2]. Also an indication and a discussion on this topic are taken in [3]. There is a downside using un-trained/naïve listeners, as they are shown to be less discriminating and less reliable compared with trained listeners when grading. However, regarding preference ratings, there are indications that the performance of un-trained listeners interested in sound quality and trained listeners is generally the same [4].

2.2 Bias in affective judgment
In their extensive literary review, Zielinski et al. describe eight biases related to affective judgments [5]. The following section presents an overview of these biases. The first two biases – appearance and branding/labelling – relate to how subjects are influenced by the appearance, brand, or the label of the equipment being tested. The third bias, expectations, refers to a subject’s expectations regarding an object under investigation

\(^1\) The descriptive method is basically a qualitative method, but the term descriptive method gives a better indication of what data the method actually collects, description of perceptions, and will be used during the course of this study.
and how well this expectation is fulfilled by the object [5]. This difference can give rise to bimodal or multimodal distribution of scores, indicating that the common practice of averaging and summarizing scores (mean values and confidence intervals) can lead to erroneous conclusions. The fourth bias presented is the non-acoustical factors of a sound, the meaning of the sound, in this paper referred to as association bias. This bias is defined by what the subject associates with the sound. For example, Zielinski et al. [5], taking an example from Fastl [6], note that German subjects and Japanese subjects have different associations for the sound of a church bell: the Germans subjects interpreted the sound of the bells as “pleasant” and “safe” and Japanese subjects interpreted it as “dangerous” and “unpleasant”. The fifth bias, the personal preference bias, essentially is the change in preference by the subjects – changes to listening habits, changes in fashion, or changes in technical quality of products [5]. The sixth and seventh bias, emotions and mood, is caused by the subject’s reaction towards a specific stimulus. If the subject is happy, the scores tend to be more positive. The mood is the bias caused by the subject’s general attitude. If the subjects have a more positive frame of mind, they tend to grade the pleasantness of a sound significantly higher than subjects with a negative attitude [5]. The eighth bias, the situational context, is the result where the product is used. Thus the sound quality might be graded differently depending on the situational context. However, it is unclear whether this bias has an effect on the listener’s evaluation of a sound [5].

2.3 Implications of affective bias
How to reduce the effects of these biases is debatable. One non-practical solution to reducing bias is to avoid tests involving affective judgments; however, this is not a good solution since affective judgment can provide useful information about market behaviour. A combination of sensory measurement and affective measurement can be applied. Thus this approach can provide additional descriptive information on products under investigation. In addition, it might be possible to increase the validity of affective judgment tests on several groups of listeners [5].

2.4 Bias involved in this research
Since an affective measurement is assumed to be composed of several factors (context, mood, emotion, background, and expectation), it is evident that they all have an effect on judgment. Subjects can be asked questions that address the following biases:

- Appearance and branding/labelling biases
- Expectation bias
- Association bias
- Personal preference bias
- Emotion and mood bias
- Context bias

This study does not address appearance and branding/labelling biases, as the subjects do not know what is under investigation. They are only instructed to rank the stimuli according to preference. In addition, the situational context bias cannot be investigated since the listening test is performed in a controlled situation and in only one context. Only the subjects’ association to different situational contexts can be investigated.
3. Framework of the descriptive research method

3.1 Foundation
As stated in the introduction, this work shifts the focus from generalization of many subjects’ perceptions to the individual subjective experience. This approach requires individuals to reflect on their experience and affective judgment. Thus this approach does not apply generalization and prediction of the results and conclusions on a selected population are therefore not available.

Previous research [7, 8,9] on perception within this framework collects data from interviews and analyses the data using a grounded theory approach, generating a theory after the data is collected. That is, the researcher does not have a predetermined theory before collecting or analysing the data rather the data suggests the theory. This approach is also frequently used in exploratory research where little is known [10]. In this research, to some extent, it is known what an affective judgment is composed of (see section 2), but not to what extent and how it is integrated in the subject’s judgment, a characteristic that makes this research project more investigative than exploratory. This limitation makes the use of grounded theory less appropriate, since there is a specific topic under investigation – the composition of affective judgment. This study relies on a phenomenological approach where experiences and how the subject perceives experiences are in focus. The essences of an experience are the core meaning/meanings that are mutually understood through a phenomenon that is commonly experienced [11]. This description of phenomenology might seem diffuse, but it essentially means that when such an approach is put into the perceptual audio evaluation context, the phenomenon under investigation is what constitutes affective judgment and the essences address how individual subjects integrate cognitive factors into affective judgment (context, expectations, previous experience, traditions, spatial impression, etc.).

3.2 Listeners
Qualitative research often uses purposeful sampling. This sampling technique is based on the assumption that the investigator wants to discover, understand, and gain insight into a phenomenon under investigation. Thus the sample selected, the subject, must have information and experience about the phenomenon [11].

3.3 Interviews
In phenomenological approaches, data typically is collected during interviews. Roulston presents a neo-positivist’s conception of interviewing, which follows phenomenology and grounded theory assumptions [12]. This conception fits with this work since it focuses on the following research questions:

- What are the participant’s beliefs, perspectives, and attitudes according to X?
- What are the participant’s experiences in relation to X?

---

2 The X refers to an experience, but could also be any phenomenon under study.
In addition, Roulston [12] identifies methodological issues a researcher should consider when relying on this particular conception:

- The interviewer takes a neutral role in the interview, does not express his/her own perspective on the research topic.
- The interviewer minimizes research influence (i.e., bias) via the use of “open” and “non-leading” questions.
- The interviewer asks “good” questions in a sequence that generates “valid” and reliable data about the topic. ([12], p 205-206)

3.4 The foundation of the phenomenological analysis.

To arrive at the essences of an experience, three steps are commonly used: bracketing, phenomenological reduction, and horizontalization. Bracketing (also called epoche) means that before starting the interviews the researcher should be aware of his/her prejudices, viewpoints, and assumptions of the phenomenon under investigation [11]. Phenomenological reduction is the process of reading and describing, looking and describing, etc. to comprehend the essences in the text and to isolate the phenomenon – understanding the text [12,13]. Phenomenological reduction and horizontalization are closely related. In horizontalization, the researcher examines and organises all the collected data (interview texts) into clusters or themes to make the essences clear. The results of this type of study are often a combined description that presents the essences of the investigated phenomenon, i.e., the structure of the experience [11]. Both phenomenological reduction and horizontalization can be done with different methods. The methods used in this research are meaning condensation, tables and graphical representation of the data, similar to mind maps. Meaning condensation, in short, is a method that condenses uttered meanings, transcribed interviews, into more essential meanings [14], condensing long explanations/descriptions into shorter manageable sentences. Meaning condensation follows the phenomenological reduction process. The tables and graphical representation were used to structure and categorize the data. See section 4.4 for a description of how the methods were used.

4. Experimental setup

4.1 Choice of Listeners

The subject selection (n=3) used the guidelines of purposeful sampling as presented in 3.2. The criteria for the subjects were as follows: naïve listeners (no previous experience with listening tests) and frequent listeners of radio or other types of media involving audio. The choice of using naïve listeners is based on the information in section 2.1. Consequently, naïve listeners (i.e. without experience from formal listening tests) with some interest or knowledge of sound and sound quality were used, as they are more likely to have the same preference as trained listeners. The use of naïve listeners would also be more likely to collect a broader motivation not directly related to sound quality parameters or characteristics. This may allow the collection of other motivators of preference other then sound quality motivators, e.g. associations to the sound or the meaning of the sound to the listener.
4.2 Stimuli selection
The stimuli used were audio files from MUSHRA [15] used in a joint experiment conducted by Swedish Radio and LTU. The stimuli were composed of speech (a female voice reading the news) and music (a pop/rock song). Each stimulus occurred in three versions; Two were perceptually coded at two different bit rates (48 kbit/s and 192 kbit/s, both using DAB+ coding) and one version was linear PCM (not perceptually coded). A total of six stimuli were used in the test. For more detailed information about how the stimuli were processed, see [15].

4.3 Listening test setup and interviews
The experiment was conducted in an acoustically controlled environment. The equipment used was a studio grade soundcard (MOTU Ultralite Mk3), near field loudspeakers (Genelec 1030A), and one laptop computer running Windows 7 and the computer program STEP³. The loudspeaker was positioned in the regular 60° position. The listening levels were adjusted according to BS.1534-1 recommendations [16]. The order of the stimuli, both content and codecs, were randomized for each subject. The randomization process was done in the STEP software session file.

Before the listening test, a short interview was conducted with each subject. The interview included questions about listening habits, expectations about the test, mood, and awareness of sound in different media. Then the subjects were instructed to rank the three versions for each stimulus, from best to worst. After each ranking (speech and music), an interview was conducted where the subject was asked to describe his/her ranking – which version was worse, best, etc. – and why they perceived them as best, worse, or in the middle. Additional questions were also asked about what method/strategy they used when conducting their ranking and if they associated the differences to something in particular. The interviewer was present during the entire test. Figure 1 provides a schematic of the test procedure. An interview guide provided the questions used during the interviews (Appendix 1).

4.4 Interview analysis
The first step in the analysis was to use meaning condensation. The sentences created in the meaning condensation process were put into tables with different topics (Table 1).

The second step of the analysis was to gain insight into how the processes of making an affective judgment are composed for each subject. The interviews were read again along with the results presented in the tables. From this analysis, a graphical representation was created that describes how each subject carried out their affective judgments. As a final step in the analysis, one single graphical representation of all the subjects’ affective judgments was created in order to show an outline of what the affective judgment was composed of. See Figure 2 for an outline of the analysis.

---

³ http://www.audio researchlabs.com
5. Results/Analysis
This section, presenting the results of the interview analysis, describes the composition of affective judgment. Interesting findings in the data are also presented.

5.1 Composition of affective judgment
Affective judgment is composed of several parts: the methods used by the subjects, the differences perceived, and how the subjects’ accomplished the ranking (Figure 3).

5.1.1 Subjects’ differentiating strategies
From the analysis depicted in Figure 3 (Method), it can be seen that the subjects have several strategies for perceiving a difference in the stimuli. To discriminate differences, the strategies include first impressions, the whole, extremes, beyond the music, and specific parts. Beyond the music has a specific meaning:

"I am also trying to listen beyond the music and the vocal; there is something in between which makes everything fit together... and this becomes often rough."

This indicates that this subject perceived a subtle quality degradation that is not related to the music but the overall quality. To discriminate differences, other subjects listened to specific parts and the proportion between the background sound and the verbal information in the stimuli (speech). One subject listened to the extremes to check for differences; when the subject became used to the extremes, the subject listened again to discriminate the differences even further.

5.1.2 Perceived differences
From the analysis, the perceived differences could also be divided into two categories, see Figure 3. These differences are based on associating the different processed stimuli with different media:

*A* (PCM): a radio commercial (earache/very clear sound/the information comes thru); *B* (48 kbps): student radio when a student that cannot handle the equipment right; and *C* (192 kbps): a kitchen radio with bad reception/windy.

The association of the differences included radio or television sound and how the particular stimuli (news or pop music) usually sound. One subject noted that at the beginning of the stimuli (speech) the sample sounded as if dynamic compression was applied to the sound and related the sound with the knowledge from the subject’s education. One subject associated each version of the music stimuli as being placed at different locations in the mix – next to the bass drum, next to the guitar player, and next to the lead singer. The direct differences included changes in frequency content between the stimuli, sharp consonants in speech, changes in level between instruments, breathing sound, and how well the listener comprehended the verbal information in the stimulus.

5.1.3 Ranking
Figure 3 (Ranking) also divides the ranking process into two categories: 1) a ranking based on a combination of the associated differences and direct differences and 2) a
ranking based on association of the differences. In the former case, it seems to be a preference ranking, by two subjects, that incorporates both difference associations and direct differences. In the latter case, one subject created two separate rankings based on associated contexts, music stimuli. One ranking based on the subject’s own preference, if the subject listened at home, and another ranking based on if the subject was doing a radio show, transmitting radio. This subject made a ranking based on the target audience for the radio. When looking at the subject’s background, information elicited from the initial interview, it can be seen that the association relates to the subject’s background since the subject works with radio and Radio Theatre, indicating that the association is intertwined with the subject’s listening habits.

The subject who ranked the stimuli two ways also ranked the speech stimuli in two ways: how quickly the quality was accepted (acceptance) and when the subject was used to the stimuli (quality). Although these rankings are not shown in Figure 2, they are evident in step 2 of the analysis. This type of ranking might indicate that there are several different ways to preference rank stimuli, making it hard for the subject to determine how to do the ranking.

5.2 Other findings

5.2.1 Mood/Expectations

From the initial interview, one clear result is that the subjects’ mood and expectation on the test was positive. All three subjects used adjectives such as curious, exciting, little nervous, interested, tired but happy, focused on doing it right, and expectant when describing the mood and expectation of the test. The adjective little nervous was interpreted as positive since it was used in the dialog along with adjectives such as curious and exciting.

5.2.2 Comparison of listening habits, education/profession, and ranking

When comparing each subjects’ answers on listening equipment and education/profession, extracted from the initial interview, with the subjects’ ranking (Table 2), one pattern emerges. Subject 2 listens to music mainly on the computer using Spotify free\(^4\) (software), and this subject also ranked the 48 kbps stimulus as most preferred because it was perceived as not as sharp as the other stimuli and had no frequency cut in the lower frequencies. This preference might indicate that the subject’s preference is low bitrate audio although further investigations are needed.

Interestingly, the subjects who studied radio were more consistent than the subject who studied TV production. This difference might be due to the experience of listening to audio that the subjects studying radio have compared to studying TV production.

\(^4\) Music software for streaming music off the Internet. It implements the format Ogg Vorbis at bitrates q3 (≈96 kbps), q5 (≈160 kbps) and q9 (≈320 kbps). The latter is a pay subscription [17].
6. Discussion

As stated, there are methodological issues regarding the interview: the interviewer does not express his/her own perspective on the research topic. To minimize the influence of the interviewer’s experience and expectations, open and non-leading questions were asked in this experiment. During the development of the interview guide, these issues were taken into consideration by creating open questions regarding the research topic and during the interview the interviewer focused on summarizing/interpreting the statements of the subject after long descriptive statements by the subjects. This strategy enabled the subjects to comment on the summary/interpretation directly during the interview, eliminating to some extent misinterpretation. During the design of the interview guide and implementation of the interviews, the researcher strove to become aware of his prejudices, viewpoints, and assumptions. Thus the interview questions focused on the subjects’ descriptions and viewpoints and not on what the researcher thought the subjects would perceive. In addition, the subject was encouraged to comment on the researcher’s summary and interpretations. These strategies helped establish an objective standpoint so as not to affect the results. This entire process can be categorized into the bracketing process.

To make the structure of experience clear, the data was read and sorted – a process called horizontalization. Horizontalization – condensation, sorting, and creation of the graphical representations of the results – was performed during the analysis. From these, the structure of the affective measurement could be seen. This process depends on the researcher’s ability to scrutinize the data and involves a high amount of interpretation in order to make phenomenon clear. To further enhance the analysis, a second researcher’s analysis could be used as a comparison, checking for similarities and differences, improving the accuracy of the interpretation.

When relating the results collected in the study to the biases in affective judgment posed by Zielinski et al. [5], there is clear evidence that the subject does incorporate them in the judgment. Association could be seen when distinguishing differences – associations related to radio commercial/student radio/radio with bad transmission reception, associating to different placements in the mix, and association to dynamic compression differences. The association could also be seen in the ranking process association to situational contexts, listening at home, or playing the stimuli on the radio. These associations could be a result imposed by the subjects trying to figure out what the differences between the stimuli are because this information was not presented to them during the experiment.

The emotion and mood bias could be checked in the initial interview since it showed the subjects’ expectations, mood, and attitudes before the experiment, a strategy that could be helpful in any experiment. There were also indications of how the personal preference bias seems to affect affective judgment when comparing the subjects’ listening habits and education with their ranking: one subject who often listened to audio with low bitrates preferred 48 kbps. This conclusion is not certain since the subject’s rating could have been influenced by his/her inexperience with the speakers used in the experiment.
The data collected contains information about the subjects’ listening habits, their initial attitudes towards the tests, what particular differences they perceived, how they evaluated the stimuli, how they associated the differences to different contexts, and what strategy they used to make an affective judgment. This information, of course, is highly dependent on the questions stated in the interviews, which depend on the research question and purpose posed. The important thing to notice from this work is that interviews and an interview analysis approach based on a phenomenological foundation can give insight into what is perceived and how the particular perceptions/judgments are composed.

One could pose the following question. Why is this method useful and why should one use it when evaluating perceptual audio? By looking at the subjects’ descriptions of their perceptions, we can collect more information on what affects the results. By implementing interview and interview analysis of this sort for a numerical experiment, a better understanding of the perception can be achieved. Since researchers have no direct access to what is going on in other human minds, the easiest way to extract information is to systematically ask the subjects to describe their experience. There is also a possibility that a combination of “regular” affective judgment tests and this descriptive approach can show what biases are or are not affecting the results. Thus the decision-making process becomes easier for the researcher on further tests. The results from this study shed light on what an affective measurement is composed of. It also indicates that interviews and interview analysis can be valid methods for understanding what perceptions are composed of.

8. Conclusions
This work shows that the descriptive method, interviews, and phenomenological analysis, where the subjects describe their perceptions freely, can shed light on what an affective judgment is composed of. In addition, the results show the methods used by the subject in order to discriminate the perceived differences, what the perceived differences are, and how the subjects incorporate background information, perceived differences, and association of the differences into their preference ranking. The results give an indication of a casual relationship between listening habits and preference ranking, but further research on this relationship is needed. The descriptive method can be fruitful when an understanding of how a perception is composed is needed.

Acknowledgements
The author would like to thank Jonas Ekeroott and Dr. Anders Persssson for their valuable comments on the manuscript.
References


8. Jumisko-Pyykkö, S; Reiter, U; Weigel, Chr, Produced Quality is Not Perceived Quality - A Qualitative Approach to Overall Audiovisual Quality; *3DTV Conference*, May 7-9, 2007


15. Percived audio quality of realistic FM and DAB+ radio broadcasting systems. Berg, J; Bustad, C; Jonsson, L; Mossberg, L; Nyberg, D. Submitted to the *Journal of Audio Engineering Society*


17. Information about Codec quality, FAQ Spotify.
URL: http://www.spotify.com/se/help/faq/tech/codec-quality/
Collected 2012-05-14
Figure captions

Figure 1: Listening test procedure. $X_1$ and $X_2$ are one of the two stimuli: speech or music.

Figure 2: Steps in the analysis

Figure 3: Composition of affective judgment

Table 1: Topics used for categorizing sentences

Table 2: Comparison between listening habits, education/profession, and ranking. PCM (Non-perceptually coded material), 192 kbps (perceptually coded material with an bit rate of 192 kbps) and 48 kbps (perceptually coded material with an bit rate of 48 kbps).

Appendix 1:
Interview guide translated from Swedish.
Figure 1:

Data

Analysis

Step 1

Transcribed interviews

Step 2

Meaning condensation

Step 3

Table presentation

Individual graphical representation

Overall Graphical representation

Figure 2:
Figure 3:

<table>
<thead>
<tr>
<th>Topics</th>
<th>Categories</th>
<th>Formats</th>
<th>Content</th>
<th>Expectations/moods/attitudes before the test and on the test.</th>
</tr>
</thead>
</table>
| Listening habits         | Time spent listening to music/audio in general | • Portable media player  
• Mobile phone  
• Computer  
• Home stereo system  
• Software used  
• Headphones  
• Loudspeakers | • Musical preference  
• What the subjects listen to. |                                                                      |
| Perceived differences    | Music                 |                                                                         | Speech                       |                                                                  |
| Ranking                  | Music                 |                                                                         | Speech                       |                                                                  |
| Association              | Music                 |                                                                         | Speech                       |                                                                  |
| Method used for looking at differences | Music |                                                                         | Speech                       |                                                                  |
Table 2:

<table>
<thead>
<tr>
<th>Subject</th>
<th>Listening equipment (listening habits)</th>
<th>Education/Professional fession</th>
<th>Ranking Speech</th>
<th>Music</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>• Mobile phone • Headphones</td>
<td>Studying TV production</td>
<td>1. PCM (C)</td>
<td>1. 192 kbps (A)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. 48 kbps (B)</td>
<td>2. PCM (B)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. 192 kbps (A)</td>
<td>3. 48 kbps (C)</td>
</tr>
<tr>
<td>2</td>
<td>• Mainly Computer and Spotify free (software) • Computer loudspeakers • Computer connected to external loudspeakers (radio) • Sometimes using a portable media player (iPod)</td>
<td>Studying radio</td>
<td>1. 48 kbps (B)</td>
<td>1. 48 kbps (C)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. 192 kbps (A)</td>
<td>2. 192 kbps (B)/PCM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. PCM (C)</td>
<td>3. (A) (hard to determine ranking order)</td>
</tr>
<tr>
<td>3</td>
<td>• Mobile phone • Radio • Computer • Computer loudspeaker • Computer connected to Hi-fi loudspeakers • iTunes (AAC files) • Sometimes Spotify</td>
<td>Studying radio, Works with; radio theatre and audio</td>
<td>Acceptance</td>
<td>Used to the stimuli</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. 192 kbps (C)</td>
<td>1. PCM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. PCM (A)</td>
<td>2. 192 kbps</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. 48 kbps (B)</td>
<td>3. 48 kbps</td>
</tr>
</tbody>
</table>

Appendix 1:

**Initial interview**

What is your favourite music?
- Give examples of music genre you prefer
- Give examples of music genre you do not prefer

How much do you spend listening to music or sounds (radio, TV, or other)?

Where do you listen?
- Radio?
- Mobile phone/portable media player?
- Internet/Spotify/iTunes?
- CD/Vinyl?
- Other media?

What are you expectations on the test?
- Are you interested?
- Is it going to be fun conducting the test?
- Do you want to “get it over with”?
- Do you feel curious? (translated from Swedish word *nyfiken*)
- Do you feel insecure?
- Do you feel uncomfortable?

What is going to happen during the test?

What is your mood?
Nyberg, Berg  A Descriptive Method for…

Nervous?
Expectant?
Happy?
Sad?
Angry?
Sulky? (translated from Swedish word sur)

Questions during the listening phase
What sound is the best according to your opinion?
   Why?

What sound is the worst according to your opinion?
   Why?

What sound is in between these two?
   Why?
   In comparison with both the best and worse, which one is closer?
Do you perceive any differences in the sound?
   Describe the differences
   Between which sounds do you perceive a difference?

Do you listen to the overall sound or are there any specific parts in which you listen to?
   Which parts?
   What is it that you listen to in particular?

Do you associate the sound to something when you compare the stimuli?
   If yes, can you describe this “something”? 

In which part of the sound do you perceive the differences?

When do the differences become apparent?

What are the characteristics of the differences?

How do you go on when comparing the sounds/stimuli?