Araştırma Makalesi

Etnomüzikoloji Dergisi
Ethnomusicology Journal
Yıl / *Year:* 7 • Sayı / *Issue:* 1
(2024)

A Methodology Proposal for the Evaluation and Ranking of 3D Sound Attributes[*]

Laçin ŞAHİN[**]
Can KARADOĞAN[***]

Abstract

This study proposes a method for evaluation and ranking attributes in 3D audio based on their functionality. According to the proposed method, a list of attributes is created from the existing research about surround and 3D sound attributes. The Semantic Differential Scale is suggested as an evaluation method. Dimensions for the semantic differential scale are offered, which are importance, comprehensibility, and noticeability. The methodology uses a weight assignment process based on expert panel opinions, minimizing individual error. The weighted average scores from the assessors that participate in the evaluation phase are combined to create a functionality score. This score is used to rank the attributes according to their functionality, which is the operationalized result of three dimensions. The study also suggests providing musical context to subjects during evaluation, allowing them to focus on the attributes. Finally, procedures such as data arrangement, rank listing, statistical and comparative analysis methods are explained.

Key words: 3D sound, 3D sound attributes, semantic differential scale, listening test, statistical analysis

[*]Makale Geliş Tarihi: 27 Mayıs 2024 Makale Kabul Tarihi: 10 Haziran 2024

[*]Produced from the doctoral thesis of the first author

[**] Doctoral Student, Istanbul Technical University, Center for Advanced Studies in Music (ITU MIAM), sahinla@itu.edu.tr https://orcid.org/0009-0005-4805-1338

[***] Professor Doctor, Istanbul Technical University, Turkish Music State Conservatoire, karadoganc@itu.edu.tr https://orcid.org/ 0000-0003-3611-6980

[*]***Acknowledgements:*** This research and related doctoral dissertation is being supported by İTÜ-BAP (Istanbul Technical University Scientific Research Projects Unit) / Project ID: 43720

3D Ses Özelliklerinin Değerlendirilmesi ve Sıralanması için Bir Metodoloji Önerisi

Özet

Bu çalışma, üç boyutlu ses alanındaki ses niteliklerini fonksiyonelliklerine göre bir değerlendirme ve sıralama yöntemi önermektedir. Bu yönteme göre alanda daha önceden yapılmış çalışmalardan bir ses nitelikleri listesi oluşturulur. Değerlendirme yöntemi olarak Semantic Differential Scale önerilir. Semantic differential scale için ise önem, anlaşılabilirlik (metinsel) ve fark edilebilirlik (duysal) şeklinde üç boyut önerilmektedir. Metodoloji, uzman görüşlerine dayalı ağırlık atama sürecini önerir. Bu şekilde bireysel hatalar en aza indirgenmiş olur. Nitelikleri değerlendirme aşamasına katılan dinleyicilerden gelen ağırlıklı ortalamalar birleştirilerek bir fonksiyonellik puanı oluşturulur. Bu puan, üç boyutlu ses niteliklerini fonksiyonelliklerine göre sıralamak amaçlı kullanılır. Bu çalışma değerlendirme sırasında katılımcılara müzikal bağlam sağlama ve böylece katılımcıların ses niteliklerine daha çok odaklanmalarına izin verme gibi prosedürler de önermektedir. Son olarak, veri düzenleme, sıralama, istatistiksel ve karşılaştırmalı analiz yöntemleri gibi prosedürler açıklanmaktadır.

Anahtar kelimeler: Üç boyutlu ses, Üç boyutlu ses nitelikleri, semantic differential scale, dinleme testi, istatistiksel analiz.

Introduction

“Rigor in the physical measurement of sound signals should be matched by equal rigor in semantics relating to subjective evaluation.” (Rumsey, 2002, p. 651). Subjective assessment of sound in audio research is a continuously evolving study area. It is concerned with the subjective evaluation of a certain sound, space, sound reproduction system, or a combination of all of them by listeners, or more properly, in the terminology of the field, subjects, or assessors. While objective assessment deals with quantifiable metrics such as sound pressure level, frequency response, signal-to-noise ratio, or total harmonic distortion, subjective assessment deals with the feelings sensed by subjects, and their responses. In the audio realm, subjective assessment research started with concert hall acoustics as early as 1900 (Sabine, 1900, as cited in Pedersen & Zacharov, 2015), later monophonic studies with loudspeakers (Gabrielsson & Sjögren, 1979), which then evolved into stereophonic studies, later into multichannel with surround systems such as 5.1 and 7.1, and finally into 3D sound systems. Although 3D audio existed for a long time in the form of multichannel audio and ambisonics, it did not garner attention from the public until recently. Likewise, attribute research in the field or 3D audio could be considered fairly new as well. One of the earliest studies about 3D audio perception was conducted by Guastavino and Katz in 2004 [Ambisonics technology was used in the study]. In their study, Guastavino and Katz (2004) summarize the sound quality assessment research [up to 2004] and define two main categories: research into monophonic reproductions of loudspeakers which is mainly concerned with timbre and distortion, and research into room acoustics which is mainly concerned with spatial attributes. The third category, the research about spatial sound perception of multi-channel audio was newly increasing [at 2004] (p. 1105). Thanks to the advent of systems such as Dolby Atmos and Auro 3D, 3D audio technology is now becoming more prevalent, and research follows it.

As previously stated, objective assessment in 3D and surround audio uses objective measurements and parameters such as: "interchannel crosstalk (ICXT), fluctuations of interaural level and time differences (ILD and ITD), interchannel correlation coefficient (ICC), interaural cross-correlation coefficient (IACC), and direct-to-reverberant energy ratio (DRR)" (Lee & Johnson, 2021, p. 871). On the other hand, subjective assessment mainly deals with attributes. Attribute is defined by ITU (2015) as: "a perceived characteristic of a hearing event, according to a given verbal or written definition" (pp. 28-30). Attributes could be seen as the verbal or written descriptors of a sound, or the feelings evoked by a sound. A listener could select stimulus A over stimulus B, but this selection does not provide a detailed information. The overall response or preference has many underlying factors. These factors are the attributes. The attributes bring multi-dimensionality to the preference of a subject and reveal the principal and secondary components. A list of attributes, focused on the spatiality of sound, from Berg and Rumsey (2001, p. 3) could be seen in Table 1.

Table 1. A list of spatial attributes by Berg and Rumsey (2001, p. 3)

General attributes
Naturalness
Presence
Preference
Envelopment
Source attributes
Source width
Localisation
Source distance
Room
Room width
Room size
Room spectral bandwidth
Room sound level
Other attributes
Background noise level

Attributes are usually selected by researchers from the literature according to their research goals and presented to listeners in a listening experiment. Listening experiments are the main tools of sound researchers when something must be tested. In addition to the testing space and the sound-reproduction system such as a multi-channel loudspeaker setup, a listening test involves sound excerpts (test stimuli), listeners (test subjects), and research questions. Subjects could be asked to evaluate a test stimulus according to certain attributes. The number of test stimuli could be many, and the test stimuli may be long, the resultant cognitive load may be exhausting, and the experiment could be time consuming. For instance, ITU (2015) recommends that a "grading phase" [listening test] session should not last more than 20-30 minutes (p. 6). Furthermore, the attributes could be too much, or inappropriate for the research goal. To avoid these problems, a researcher should always aim for an experiment that is goal-oriented, not tiresome and not time or money consuming. This could be easily accomplished by limiting the test stimuli in terms of amount, excerpt duration and total duration. Besides this, a researcher could limit the number of attributes to be rated. By limiting the number of attributes, the duration of the listening experiment could be reduced, the focus of the subjects increased, and their cognitive load decreased. In short, the effectiveness of the test could be enhanced.

There are a vast number of attributes and related research in stereophonic, surround, and 3D sound studies. Some studies are concerned with the discovery and elicitation of these attributes; some are interested in the categorization of these attributes (Berg & Rumsey, 2001; Le Bagousse et al., 2010b; Pedersen, 2008); some are interested in elicitation and categorization of these attributes (Koivuniemi & Zacharov, 2001; Guastavino & Katz, 2004); some are interested in the underlying objective measurements; some are interested in the clustering, reduction, and selection of these attributes (Francombe et al., 2017); and some studies focus on a single attribute (Mason, 2017). Although there are researchers that spend a certain amount of time eliciting these attributes from expert assessors before their actual experiment or listening test stage, most experimenters select a set of attributes from previous research that are appropriate for research goals and conduct their experiments by using these attributes. There is no guideline or method for the selection of these attributes. Moreover, the selection stage could be confusing for the experimenter. A list of attributes carefully selected by expert assessors for a certain research field could be beneficial for researchers. Therefore, the main goal of this study is to propose a research method for selecting and ranking attributes in the context of 3D audio in terms of their significance and functionality.

Experiment Design / Methodology

Preselection of Attributes

The proposed procedure of this study forgoes the elicitation method, where a group of expert assessors are given the task of coming up with their own words, descriptors, adjectives, and attributes. This method has been implemented in many previous studies, and there is a body of work that has conceived, proposed, and defined a large number of attributes that could be used in the context of 3D sound. For this reason, the proposed method of this study is to make a scientific selection from this list. The first thing that comes to mind is to list all these attributes and submit them to a panel of expert assessors for selection, reduction, and ranking. However, judging by the number of available attributes in the literature, this process would be time-consuming, which would defeat the purpose of this study: time efficiency. Thus, a preselection phase before the actual selection and ranking is necessary.

The preselection could be done by the experimenters themselves by grouping attributes together according to the similarity of their definitions, concurrence in literature, and relevance to the field which is 3D sound. The main objective in this phase should be to limit the list of attributes that are going to be submitted to the panel of experts. The preselection phase could limit the number of attributes to be evaluated under 100, or more conveniently, around 50. ITU (2015) points out that, a grading session should not exceed 30 minutes (p. 6). If we basically assume that an average assessor spends 30 seconds for the evaluation of 1 attribute, the total duration will result in 25 minutes, which is around the suggested duration for listening experiments by ITU. However, if the assessors are reading multiple definitions of these attributes from different authors, the total duration would increase and the total experiment duration would exceed the suggested durations of ITU. In short, a preselection phase could reduce the number of attributes to 50 before the actual selection and ranking phase.

Selection and Ranking of Attributes

How should assessors select and rank 50 attributes? What should be the number of attributes in the final list? What should be the selection criteria? Firstly, the most basic approach would be to ask the assessors to select a number of attributes according to their importance for the field of 3D audio. A similar approach would be to ask the assessors to rank the 50 attributes according to their importance. Another method would be to introduce an importance scale such as 1-100 or 1-5 and ask the assessors to score the importance of 50 attributes. Next, the experimenter could calculate the average scores of all attributes and rank them accordingly. However, this approach contains the same problem that the common method of preference selection presents: the underlying factors are hidden.

Rating Scale: Semantic Differential Scale

A more advanced method would be to propose a set of criteria according to which listeners can evaluate the attributes, in other words, attributes of attributes. The proposed method that involves this type of evaluation design is the "Semantic

Differential Scale" (Osgood et al., 1957). Robson and McCartan explain that rather than measuring their [assessors'] level of belief in a specific idea, it [semantic differential scale] focuses on evaluating the subjective interpretation of a notion by the responder (2016, p. 314). Semantic differential scale allows the assessors to evaluate the test items on different dimensions. Dimensions are presented as "…a series of bipolar rating scales…" (Robson & McCartan, 2016, p. 314). Or, it could be said that when adjectives are presented on a horizontal axis with end-words they become "dimensions". For example, the attribute "clarity" could become a dimension when it is presented with anchor points such as "not clear" and "very clear". This multi-dimensional aspect of the semantic differential scale allows us to pinpoint the underlying contributors to an otherwise obscure preference. In fact, Cozby and Bates state that "semantic differential scale is a measure of the meaning of concepts" (2018, p. 240). There is no limit to the number of dimensions to be used, but again, time efficiency for the final test should be of concern, and a limit of 3 or 4 dimensions could be aimed at. Table 2. shows the semantic differential scale used for an imaginary test question that asks subjects to rate a set of headphones on three dimensions on a scale of 5.

Table 2. Semantic differential scale used for a headphone evaluation experiment

Question: Please rate the set of headphones according to its comfort, bass response, and noise-cancellation
Very uncomfortable ______ ______ ______ Very comfortable
Very inadequate bass ______ ______ ______ Very adequate bass
Very bad cancellation ______ ______ ______ Very good cancellation

Table 3 shows the arbitrary responses of an imaginary test subject:

Table 3. Responses given for a headphone evaluation experiment using the semantic differential scale

Very uncomfortable ______ __X__ ______ Very comfortable

Very inadequate bass ______ ______ ______ [✓ Very adequate bass]

Very bad cancellation __X__ ______ ______ Very good cancellation

The responses given by the test subject on Table 3. should be interpreted as follows: the test subject has given a score of 3 for comfort because the selection mark is in the middle; the test subject has given a score of 5 for bass response because they circled the right end-word; and finally, the test subject has given a score of 2 for noise-cancellation. The scale used in this example is a 5-point scale. However, depending on the nature of the experiment, other scales could be used as well (Robson & McCartan, 2016, p. 314). In one of the earliest multi-channel listening tests a 7-Point Likert scale was used by Nakayama et al. (1971). EBU (2000, p. 4) suggests a 6-point evaluation scale {rank scale instead of continuous -more on this on later chapters}.

Semantic differential scale is not uncommon in attribute studies. In fact, among others, specifically for the case of 3D audio assessment, it was used by Hamasaki et al. in their 2006 study about 22.2 loudspeaker system, and also by Shim et al. (2010) to compare 22.2, 10.2 and 5.1 systems.

Dimension Selection

The dimensions to be used in the case of 3D sound attributes should be carefully selected. They should reflect different aspects of the attributes and contribute to their functionality in different ways. According to Cozby and Bates, for semantic differential scale, the concepts are generally assessed based on three fundamental dimensions: evaluation, activity, and potency (2018, pp. 240-241). Activity pertains to the degree of association between the concept and action, evaluation is related to the general positive impression linked to it, and potency corresponds to its total power or "importance" (Robson & McCartan, 2016, p. 314). In fact, these are the dimension groups that the actual dimensions are usually clustered into. About dimension creation, Robson and McCartan indicate that a compilation of suitable adjective pairs is {could be} prepared for the specific notion you are attempting to assess (2016, p. 314). In light of these, the proposed dimensions for 3D sound attribute evaluation are shown in Table 4.

Table 4. Proposed dimensions

Importance
Comprehensibility
Noticeability

Here is how the dimensions would look with a 5-point scale and endpoints (Table 5.):

Table 5. Scale as it would be used in the research

Unimportant ______ ______ ______ Very Important

Not comprehensible ______ ______ ______ Easily comprehensible

Not noticeable ______ ______ ______ Very noticeable

The importance dimension, which is also used in the original concept of semantic differential scale as is, calculates the significance of the attribute in the context of 3D audio. Some attributes might be more important in other contexts, such as concert hall acoustics, loudspeakers, timbral studies etc., whereas other attributes might be more important in the context of 3D audio. As a result, incorporating "importance" as a dimension would serve the purpose of testing the significance of an attribute in the context of 3D audio. The explanatory question for the importance dimension would be: "How important do you think this attribute is in the context of 3D audio?". The end-words would be "unimportant" and "very important".

The comprehensibility dimension investigates the clarity of the definition, wording, or description of the attribute for the assessor. EBU declares that: "In all subjective evaluations there is a danger that the assessment will be unreliable because different listeners may put different interpretations on the parameters. Experience has shown that this leads to overlaps in the scores given to different parameters" (1997, p. 8). Challenge in the accurate definition of the attributes and making sure that all assessors have a similar understanding of the attributes to avoid bias in the results was emphasized by Le Bagousse et al. as well (2010a, p. 2). Some attributes are easy to understand, while others are hard to understand. Some have clear and basic definitions, and some have long, sophisticated definitions. When presented with an attribute in an experiment, a subject would like to know its meaning. Generally, these definitions are presented to the subjects before the test in a written or verbal format. The expertise of the subjects comes into play in this situation. If an attribute's definition is quickly and correctly understood, the evaluation task becomes more efficient. However, if the definition is not easily understood or wrongfully understood, the task would be inefficient or might yield the wrong results. Therefore, the comprehensibility of an attribute is a contributor to its overall functionality. The explanatory question for the comprehensibility dimension would be: "How easily comprehensible is the definition of this attribute?". Indeed, another question could also be added: "How coherent are the meaning and the name of the attribute?". The end-words would be "not comprehensible" and "easily comprehensible".

The noticeability dimension is the auditory version of the comprehensibility dimension. It examines how easy it is to notice the attribute aurally. An attribute might be important and easy to comprehend, but hard to notice by ear, or vice versa. This could affect the functionality of an attribute. If no one can hear or perceive the attribute, then could it still be important? Or does it matter if its definition is easily understood? In a 2005 study by Lee and Rumsey about the effect of interchannel crosstalk in multichannel microphone techniques, participants were instructed to assess the level of audibility of the chosen attributes. This was done to assess the relative importance of those traits and thereby decrease the number of attribute scales that need to be evaluated (2005, p. 3). There are many attributes in the literature that listeners struggle to perceive. Taking this into consideration, the noticeability dimension could be an important factor that could affect the functionality of an attribute. The explanatory question for the noticeability dimension would be: “How easy is it for you to notice the presence or absence of this attribute aurally?”. Here, the “absence” should also be presented because the absence of an attribute would also contribute to its noticeability. The end-words would be “not noticeable” and “very noticeable”.

Average Score, Weighted Average Score, Functionality Score

The weight is the affecting power of a variable. In mathematical terms, it is the coefficient of a variable. In Table 6., imaginary scores given by a test subject to the attribute “envelopment” have been presented using the previously mentioned dimensions: importance, comprehensibility and noticeability.

Table 6. Imaginary scores given by a test subject for the “Envelopment” attribute

Unimportant _____ _____ __X__ Very Important

Not comprehensible _____ _____ _____ [✓ Easily comprehensible]

Not noticeable __X__ _____ _____ Very noticeable

Here, the test subject has given a score of 4 for the importance dimension, a score of 5 for the comprehensibility dimension, and a score of 2 for the noticeability dimension. Robson and McCartan state that the scores could be summed or averaged (2016, p. 314). If we calculate the average of the three scores, the result would be 3.66 {rounded down to two decimals}.

$$(4 + 5 + 2) / 3 = 3.66$$

We can label this final score as the “functionality score”. In this calculation, each dimension has equal weight in the functionality score. In other words, each dimension has an equal influence or effect on the functionality score. They each have a %33.3 contribution to the final score, or they each have a coefficient of 0.33 {rounded down to two decimals}. However, it is the judgment and decision of the experimenter to treat them as equal. In reality, they might have different weights and different impacts on the final functionality score. If we assume that comprehensibility is less influential in the functionality of an attribute and importance and noticeability are more influential, we can assign different weights, which would give us a different functionality score. For example, we could assign the weights as follows (Table 7.):

Table 7. Arbitrary weight assignment

Importance: 0.40 (or %40)
Comprehensibility: 0.20 (or %20)
Noticeability: 0.40 (or %40)
(The total should be %100)

If we take the same scores from the previous example (Table 6.) and do the calculation again according to the weights from Table 7. we will have a different functionality score (Table 8.):

Table 8. New functionality score using the weighted averages

Unimportant _____ _____ __X__ Very Important

Not comprehensible _____ _____ _____ [✓ Easily comprehensible]

Not noticeable __X__ _____ _____ Very noticeable

Importance: 4

Comprehensibility: 5

Noticeability: 2

4 x 0.4 (Score x Weight of importance) = 1.6

5 x 0.2 (Score x Weight of comprehensibility) = 1.0

2 x 0.4 (Score x Weight of importance) = 0.8

New functionality score = 1.6 + 1.0 + 0.8 = 3.4

Previous functionality score = 3.66

Difference = - 0.26

As we can see from the new calculation, there is a -0.26 difference between the previous functionality score and the new functionality score using the new weights. In the previous calculation, the comprehensibility score of 5 had a coefficient of 0.33 (equal weight) and played a significant role in the final score. However, in the new calculation with the new weights, the comprehensibility dimension was deemed less important, and even though the given score is 5, with the new coefficient of 0.2, the power or effect of comprehensibility is reduced. This implies that weight assignment is critical in the calculation and formulation of the final scores and the resulting rank order of attributes. In the following section, the proposed approach for weight assignment is described.

Weight Assignment Process - Test Phase I

Rather than assigning the weights based on the subjective opinion of the experimenter(s), it is proposed by this methodology to assign the weights based on the collective opinions of an expert panel. This way, the error in the judgment of a single individual would be dispersed over a group of individuals and minimized.

It is proposed by this methodology to design the weight assignment phase as the first phase of the experimental design. Firstly, a group of expert assessors from the field of sound engineering, and more specifically, from the field of 3D audio should be formed by the experimenter. Secondly, the experts should be informed about the final phase of the experiment, which is the reduction, selection, and ranking of the attributes according to their functionality scores. Thirdly, the experts should be asked to assess the effectiveness, influence, or power of each dimension on the final score. In mathematical terms, they should assign the percentage or the coefficients of the dimensions. Finally, the mean scores of the individual assessments should be turned into the final weights of the dimensions.

Evaluating the Attributes According to Semantic Differential Scale - Phase II

After the initial phase of weight assignment, another group of experts should be gathered by the experimenter for the actual evaluation of the attributes. This group should also be composed of expert assessors experienced in the field of sound engineering, and again, in particular, the field of 3D audio. Firstly, the subjects should be informed about the experimental goals. They should be informed about the semantic differential scale and the three dimensions of importance, comprehensibility, and noticeability. Secondly, they should be presented with a list of preselected 3D sound attributes. The list should include the definitions of the attributes, preferably from more than one author in cases of diverse or contrasting definitions. Some attributes might include visual definitions as well, and these visual definitions should also be provided. Finally, the test scores of each subject should be turned into quantifiable data. These stages will be detailed in the following sections.

Experiment Conditions

Evaluation in Context - Listening Test

Another important suggestion of this methodology is to provide the subjects with musical context. Rather than treating the field of 3D sound and 3D sound attributes as a whole, it is recommended to focus on a specific subject under this broad area.

3D sound may include subjects such as multichannel loudspeaker systems (Zacharov et al., 2016), binaural (Qiao et al., 2022), VR and ambisonics (Millns & Lee, 2018), object-based audio, channel-based audio, 3D microphone techniques (Howie et al., 2018; Lee & Johnson, 2019), 3D production, etc. An attribute conceived for one of these subjects may not function or may not be as important in another. For this reason, selecting a specific subject is crucial. The next level of limitation would be to focus on a particular type of music or sound. EBU (2000) claims that evaluating diverse genres of program content in a single listening session is challenging (p. 5). This could mean that focusing on a single genre, or subject area in listening tests could be more effective. For example: if the experiment is about the 3D sound attributes used in 3D music productions, then selecting a specific genre would be effective. After that, it is the suggestion of this proposal to the experimenter to provide a 3D sound excerpt or a selection of excerpts to the assessors during the second phase, where the subjects are evaluating the attributes.

A survey of the literature about research into the field of sound attributes, or more specifically, 3D sound attributes reveals that there are two types of studies: studies that provide a listening test, and studies that do not. Studies that do provide a listening test generally incorporate a selection of music from different genres or provide different types of productions from a specific genre and let subjects differentiate between them by using the provided attributes. Choosing a specific genre could also minimize the assessor bias and preference. In their study, Choisel and Wickelmaier (2007) found that the kind of program material had a substantial impact on attributes and overall preference, indicating that the perceptual effects caused by the chosen reproduction modes rely on the program material to which they are applied to (p. 398). Henceforth, the suggested method here is to provide a selection of excerpts from a specific genre, or more specifically, a selection of mix sessions for the subjects. However, since the goal is to rate the attributes and not the excerpts, subjects should be warned not to compare the excerpts, and rather switch between them if necessary. For example, if the experiment is about the sound attributes in 3D pop music mixes, then it would be best to provide the subjects with a readily available 3D pop mixing session(s). This way, while evaluating the attributes, subjects would be able to listen to the context in which the attributes are asked to be used. Then, the subjects would be able to delve into different aspects of mixing and listen to possible scenarios where the attributes might become more important or more noticeable. To support this further, the subjects would be directed to take part in the test one by one. Lee (2012, p. 3) asserts that a listener's spatial perceptions could change based on their position in the listening space. By directing the participants to take part in the evaluation test one by one, every assessor can sit in the critical listening position of the speaker setup. In short, goal-specific / genre-specific test stimuli or listening material should be provided to the assessors. More specifically, this listening material should be created by the experimenter for the purpose of this test.

Listening Material - 3D Sound Recording

The listening material should match the 3D research subject of the experimenter for which the attributes are going to be evaluated for. For example: if the research subject is 3D pop productions with object-based audio, then a 3D pop mixing session that incorporates mixing elements with objects should be provided to the test subjects. If the research subject is about 3D chamber music productions with channel-based audio, then a 3D chamber music mixing session should be provided to the test subject. In fact, a recording should be done using 3D microphone techniques such as PCMA-3D (Lee & Gribben, 2014; Lee & Johnson, 2021) or OCT-3D (Theile & Wittek, 2011). Since the subjects are being provided with a listening material, optimal listening conditions should be provided to the subjects, and if the experiment is about 3D sound productions reproduced using 3D loudspeaker setups, then assessors should be seated in the critical listening position of the 3D speaker setup, and they should complete the test one by one.

Listening Test Space and Calibration

If the research is about binaural 3D renders or similar topics such as HRTF curves and profiles, the listening test space should provide the assessors with suitable headphones and adequate listening conditions. If the research is about loudspeaker reproductions, then the listening test space should provide the loudspeaker setup and room conditions of standards, guidelines and recommendations such as from ITU (2015) or ITU (2022). If these conditions are not met, they should be specified. A standard for a multichannel loudspeaker setup [with height channels] from ITU (2015) could be seen in Figure 2.

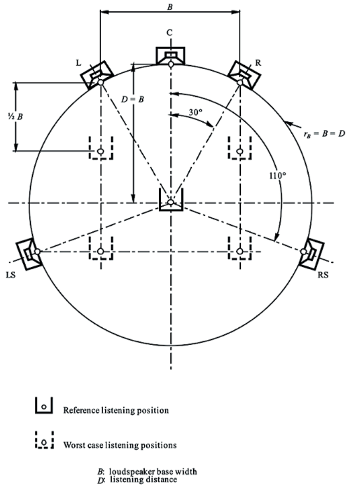
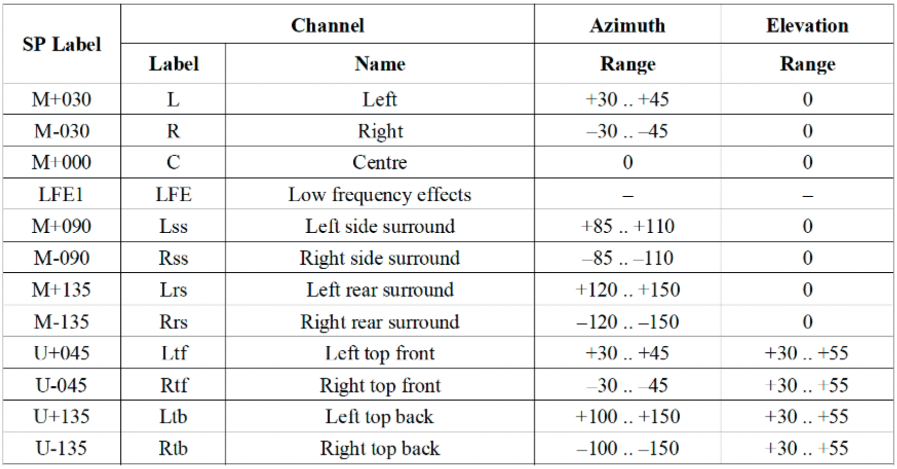

B
C
L
R
D = B
½ B
30°
$r_B = B = D$
110°
LS
RS
Reference listening position
Worst case listening positions
B: loudspeaker base width
D: listening distance
BS.1116-07

Figure 2. "Test listening arrangement with loudspeakers L/C/R and LS/RS for multichannel sound systems with small impairments" (ITU, 2015, p. 20).

In 2022, ITU revised its reccommendations with the inclusion of Advanced Sound Systems, which describes the multichannel loudspeaker setups with height [and bottom] layers that reproduce object and channe-based signals. A loudspeaker configuration recommendation for 4+7+0, which is the grouping for 7.1.4 could be seen in Table 9.

Table 9. "Loudspeaker configuration for Sound System J (4+7+0)" (ITU, 2022, p. 15)



SP Label	Channel		Azimuth	Elevation
	Label	Name	Range	Range
M+030	L	Left	+30 .. +45	0
M-030	R	Right	–30 .. –45	0
M+000	C	Centre	0	0
LFE1	LFE	Low frequency effects	–	–
M+090	Lss	Left side surround	+85 .. +110	0
M-090	Rss	Right side surround	–85 .. –110	0
M+135	Lrs	Left rear surround	+120 .. +150	0
M-135	Rrs	Right rear surround	–120 .. –150	0
U+045	Ltf	Left top front	+30 .. +45	+30 .. +55
U-045	Rtf	Right top front	–30 .. –45	+30 .. +55
U+135	Ltb	Left top back	+100 .. +150	+30 .. +55
U-135	Rtb	Right top back	–100 .. –150	+30 .. +55

Calibration of the listening rooms according to standards and recommendations is highly important for listening tests. EBU (2000) recommends a reference listening level of 78 dB SPL per loudspeaker [for surround systems], (p. 4). Most listening tests use a dummy-head or a calibration microphone to calibrate the sound pressure levels to level recommendations suggests several listening test levels, however, these levels might not be kept consistent in a scenario suggested by this methodology where assessors are free to bring the faders up and down in the mixing session, which would change the levels all the time. It is unclear whether this would have any effect on the answers provided by the assessors. Nevertheless, this information should be provided to the assessors, and readers of the research.

Subjects / Listening Panel

Expert Panel and Number of Subjects

Expertise and number of subjects are directly related topics. ITU (1997; also in 2015) reports that when the parameters of a listening test are carefully regulated in terms of both technical and behavioral aspects, […] it is frequently enough to have data from 20 participants in order to make accurate conclusions from the test (p. 3). Selecting a limited number of experts falls under the concept of purposive sampling. According to Cozby and Bates (2018), the aim of purposive sampling is to: "obtain a sample of people who meet some predetermined criterion… This is a good way to limit the sample to a certain group of people" (p. 261). The criteria could be that subjects should be experts in the topic of the experiment. In fact, many sources suggest selecting a panel or group or expert listeners or assessors for subjective listening and evaluation tests (EBU, 2000; ITU, 2015). However, the results gathered by using expert subjects would be impractical if they cannot be applied to general public. Regarding this matter, Howie et al. claim that "using experienced, trained or practiced listeners for audio evaluation requires fewer subjects and time compared to naive listeners, but should yield results that can be generalized to a larger population." (2019, p. 783). From the potential subjects, the experimenter could necessitate an experience in the field of 3D audio and a degree in the relevant field such as sound engineering. These features could work as criteria to identify subjects as experts or non-experts. Finally, the demographics data such as age (For example: Howie et al., 2018) and gender could be gathered and provided in the test results as well.

Hearing / Critical Listening

Some listening tests calculate the critical listening skills of the test subjects before the experiment. Test subjects that fail to meet the requirements of the experimenters are left out of the experiment. Some listening tests require the subjects to have an "ontologically normal hearing" according to ISO Standard 389, otherwise they are not treated as expert listeners (EBU, 2000). On the other hand, some listening tests only ask the subjects whether they have any hearing problems or impairments. However, they do not test these conditions and only provide the data.

Statistical Analysis

Determining the Type of Data

At the end of the 3D sound attribute evaluation test, the data should be gathered and evaluated statistically. About the goal of statistical analysis, ITU reports:

> The fundamental aim of the statistical analysis of test results is to identify accurately the average performance of each of the systems under test and the reliability of any differences among those average performance figures. The latter aspect requires estimation of the variability or variance of the results. (ITU, 2015, p. 21)

In a statistical analysis, the type of analysis should match the type of data and scale. Nominal and ordinal scales are analyzed using non-parametric tests, and interval and ratio scales are analyzed using parametric tests. A data or numerical value produced from semantic differential scale could be treated as ordinal data since the distances between each rank is not known or would not be equal. Cozby and Bates (2018) give the examples of letter grades and movie rating systems (p. 191) for ordinal scale. The proposed method for this study suggests converting the markings and circling on the semantic differential scale to numeric data. If the scale is a 5-point scale, the end-words should be converted to 1 and 5, and the middle positions should be converted to 2, 3 or 4, depending on the marking of the assessor. This conversion should be done for all the assessors, and the scores should be compiled.

Robson and McCartan (2016) warn that this conversion from ordinal values to scores with equal distances could create problems, however, they also mention that it should not prevent an experimenter from applying fundamental statistical analysis, because it would probably illuminate the meaning conveyed by the data (p. 416). On the contrary, Bech and Zacharov (2006) identify the 5-point Likert scales such as from ITU-R. Recommendation BS.1116-1 (ITU, 1997) as an interval scale (p. 71). In fact, Bech and Zacharov assert that [citing Nunnally & Bernstein, 1994] scales with 11 or more categories [or ranks] could become continuous scales (2006, p. 72). The notion here is that, if there are enough categories, the scale could resemble a continuous scale. Regardless, it could be said that treating 5-point Likert scale as ordinal or interval scale is debatable, and authors have different opinions on the matter. Bech and Zacharov offer a practical approach to this uncertainty. Bech and Zacharov suggest that a data presumed as interval could be analyzed using quantitative methods [parametric tests] unless it violates the statistical assumptions [such as normality], otherwise categorical methods [non-parametric tests] should be applied (2006, p. 71). This would indicate that, conversion from ranks/categories such as “not important” and “very important” to numbers 1 to 5 is non-problematic on its own. The crucial point is to check if the data set meets the statistical assumptions of parametric tests, and if not opt for non-parametric tests. The initial step to check for violations is testing the normality of the data, which can be achieved through Shapiro-Wilk’s test of normality. But before that, rank orders according to the following calculations should be created (Table 10.):

Table 10. Rank orders

1	Rank order according to average scores
2	Rank order according to weighted average scores
3	Rank order according to the average scores of importance dimension
4	Rank order according to the average scores of comprehensibility dimension
5	Rank order according to the average scores of noticeability dimension

Parametric and Non-Parametric Tests

If the Shapiro-Wilk's test of normality reveals that the data shows normality, parametric tests such as t-test could be applied. If the data shows a deviation from normality, non-parametric statistical analysis methods like the Wilcoxon Signed-Rank test, should be applied [Also suggested by EBU, 1997, for pairs of dependent variables}. In their statistical analysis section of their 2018 study about 360° microphone techniques for virtual reality, Millns and Lee employed the Shapiro-Wilks' test of normality which revealed non-normal distribution of listening test data. Thus, researchers used non-parametric tests such as Friedman and Wilcoxon signed-rank tests to analyze the data (p. 5). Both the t-test and the Wilcoxon Signed-Rank test evaluates whether there is a statistically significant difference between the two sets of data. The t-test looks for the differences between the means and Wilcoxon Signed-Rank test looks for the difference between the medians. A prerequisite for these tests is that the two sets must be a matched pair, or paired samples. In this context, attributes serve as matched pairs, remaining constant while their scores vary based on different calculations or conditions. For instance, scores could be derived from averages (with equal weight assigned to dimensions), or from weighted averages, where weights are determined from Phase I. An imaginary test result with 4 attributes' scores for two conditions could be seen in Table 11.

Table 11. Imaginary test results with 4 attributes

Attribute	Condition 1 (Average Scores)	Condition 2 (Weighted Average Scores)
Clarity	3.0	2.8
Spaciousness	3.66	3.8
Localization	4.66	4.8
Realism	4	4.2

The Wilcoxon Signed-Rank test, which is the non-parametric test used for paired samples, the compares the median of differences and rank orders of the two conditions, determining if a statistically significant difference exists between them. If there is a statistically significant difference, then it could be argued that the experiment has produced results that are statistically different from:

1- Rank order of the attributes according to 1 experimenter, only considering the importance
2- Rank order of the attributes according to 20 assessors, only considering the importance
3- Rank order of the attributes according to 20 assessors, using a semantic differential scale with three dimensions with equal weights
4- Rank order of the attributes according to 20 assessors, using a semantic differential scale with three dimensions with weights assigned by 1 experimenter.

If the data is not normal, and if there are more than two groups of matched pairs, then Friedman test could be used as well. From Table 10., it could be seen that there are 5 possible rank orders, or categories that the attribute scores could be calculated for. For example, Friedman test could be used to look for differences between average, or weighted average scores for importance, comprehensibility, and noticeability.

Finally, the top 10 or 20 attributes from all the rankings could be investigated in terms of similarities and differences. Moreover, it should be noted if there are a num-ber of attributes that could be grouped into categories or clusters. To conclude, the top 10 attributes could be considered for further and more detailed research.

Conclusion

In this study, a methodology has been proposed to select and rank 3D sound attributes. Stages such as preselection and reduction of attributes, weight assignment for dimensions, and attribute evaluation using the semantic differential scale have been described. Furthermore, in-context attribute evaluation with readily available listening material has been highlighted. In addition, the ways to record this material, and reproduce it in suitable listening conditions that adhere to recommendations were reported. Finally, data arrangement, ensuant statistical analysis, and investigation of the test results have been explained.

References

Bech, S. & Zacharov, N. (2006). Perceptual audio evaluation: theory, method and application. John Wiley & Sons.

Berg, J., & Rumsey, F. (2001, June 21-24). Verification and correlation of attributes

used for describing the spatial quality of reproduced sound {Paper presentation}. Audio Engineering Society 19th International Conference, Schloss Elmau, Germany.

Choisel, S., & Wickelmaier, F. (2007). Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. Journal of the Acoustical Society of America, 121(1), 388-400.

Cozby, P. C., & Bates, S. C. (2018). Methods in behavioral research (13th ed.). McGraw Hill Education

European Broadcasting Union (1997). Assessment methods for the subjective evaluation of the quality of sound programme material – Music (EBU Tech 3286-E). https://www.ebu.ch/home

European Broadcasting Union (2000). Assessment methods for the subjective evaluation of the quality of sound programme material – Supplement 1 - Multichannel (EBU Tech 3286-E). https://www.ebu.ch/home

Francombe, J., Brookes, T., & Mason, R. (2017). Evaluation of spatial audio reproduction methods (part 1): Elicitation of perceptual differences. Journal of the Audio Engineering Society, 65(3), 198-211.

Gabrielsson, A., & Sjögren, H. (1979). Perceived sound quality of sound-reproducing systems. Journal of the Acoustical Society of America, 65(4), 1019–1033.

Guastavino, C., & Katz, B. F. G. (2004). Perceptual evaluation of multi-dimensional spatial audio reproduction. Journal of the Acoustical Society of America, 116(2), 1105–1115.

Hamasaki K., Nishiguchi, T., Hiyama K., & Okumura R. (2006, May 20-23). Effectiveness of height information for reproducing presence and reality in multichannel audio system [Paper presentation]. Audio Engineering Society 120th Convention, Paris, France.

Howie, W., Martin, D., Benson, D. H., Kelly, J., & King, R. (2018, August 7-9). Subjective and objective evaluation of 9ch three-dimensional acoustic music recording techniques [Paper presentation]. AES International Conference on Spatial Reproduction - Aesthetics and Science, Tokyo, Japan.

Howie, W., Martin, D., Kim, S., Kamekawa, T., & King, R. (2019). Effect of audio production experience, musical training, and age on listener performance in 3D audio evaluation. Journal of the Audio Engineering Society, 67(10), 782–794.

International Telecommunication Union (1997). Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems (ITU-R BS.1116-1) https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1116-1-199710-S!!PDF-E.pdf

International Telecommunication Union (2015). Methods for the subjective assessment of small impairments in audio systems: broadcasting service (sound) (ITU-R BS.1116-3). https://www.itu.int/rec/R-REC-BS.1116-3-201502-I/en

International Telecommunication Union (2022). Advanced sound system for programme production: broadcasting service (sound) (ITU-R BS.2051-3). https://www.itu.int/rec/R-REC-BS.2051-3-202205-I/en

Koivuniemi, K., & Zacharov, N. (2001, September 21-24). Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training [Paper presentation]. Audio Engineering Society 111th Convention, New York, NY, United States.

Le Bagousse, S., Colomes, C., & Paquier, M. (2010a, June 13-15). State of the art on subjective assessment of spatial sound quality [Paper presentation]. Audio Engineering Society 38th International Conference, Piteå, Sweden.

Le Bagousse, S., Paquier, M., & Colomes, C. (2010b, November 4-7). Families of sound attributes for assessment of spatial audio [Paper presentation]. Audio Engineering Society 129th Convention, San Francisco, CA, United States.

Lee, H. (2012, April 26-29). Subjective evaluations of perspective control microphone array (PCMA) [Paper presentation]. Audio Engineering Society 132nd Convention, Budapest, Hungary.

Lee, H., & Gribben, C. (2014). Effect of vertical microphone layer spacing for a 3D microphone array. Journal of the Audio Engineering Society, 62(12), 870-884.

Lee, H., & Johnson, D. (2019, October 16-19). An open-access database of 3D microphone array recordings [Paper presentation]. Audio Engineering Society 147th Convention, New York, USA.

Lee, H., & Johnson, D. (2021). 3D microphone array comparison: objective measurements. Journal of the Audio Engineering Society, 69(11), 871-887.

Lee, H., & Rumsey, R. (2005, May 28-31). Investigation into the effect of interchannel crosstalk in multichannel microphone technique [Paper presentation]. Audio Engineering Society 118th Convention, Barcelona, Spain.

Mason, R. (2017, May 20-23). How important is accurate localisation in reproduced sound? [Paper presentation]. Audio Engineering Society 142nd Convention, Berlin, Germany.

Millns, C., & Lee, H. (2018, May 23-26). An investigation into spatial attributes of 360° microphone techniques for virtual reality [Paper presentation]. Audio Engineering Society 144th Convention, Milan, Italy.

Nakayama, T., Miura, T., Kosaka, O., Okamoto, M., & Shiga, T. (1971). Subjective assessment of multichannel reproduction. Journal of the Audio Engineering Society, 19(9), 744–751.

Nunnally, J. C. & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. University of Illinois Press

Pedersen, T., H. (2008). The semantic space of sounds. DELTA.

Pedersen, T. H., & Zacharov, N. (2015, May 7-10). The development of a sound wheel for reproduced sound [Paper presentation]. Audio Engineering Society 138th Convention, Warsaw, Poland

Qiao, Y., Zacharov, N., & Hoffmann, P. F. (2022). Prediction of timbral, spatial, and overall audio quality with independent auditory feature mapping [Paper presentation]. Audio Engineering Society 153rd Convention

Robson, C., & McCartan, K. (2016). Real world research (4th ed.). Wiley

Rumsey, F. (2002). Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm. Journal of the Audio Engineering Society, 50(9), 651–666.

Sabine, W. C. (1922). Collected papers on acoustics. Harvard University Press. (Original work published 1900).

Shim, H. Oh, E., Ko, S., & Park, S. H. (2010, November 4-7). Perceptual evaluation of spatial audio quality {Paper presentation}. Audio Engineering Society 129th Convention, San Francisco, CA, USA

Theile, G. & Wittek, H. (2011, May 13-16). Principles in surround recordings with height {Paper presentation}. Audio Engineering Society 130th Convention, London, UK

Zacharov, N., Pike, C., Melchior, F. & Worch, T. (2016). Next generation audio system assessment using the multiple stimulus ideal profile method {Paper presentation}. 8th International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal.