ACOUSTICS 2012

Assessment of spatial audio quality based on sound attributes

S. Le Bagousse[a], M. Paquier[b] and C. Colomes[a]

[a]Orange Labs, 4 rue du clos courtel, 35510 Cesson Sevigné, France
[b]Université de Bretagne Occidentale, 6, avenue Victor Le Gorgeu, CS 93837, 29238 Brest Cedex 3, France
sarah.lebagousse@orange.com

Spatial audio technologies become very important in audio broadcast services. But, there is a lack of methods for evaluating spatial audio quality. Standards do not take into account spatial dimension of sound and assessments are limited to the overall quality particularly in the context of audio coding. Through different elicitation methods, a long list of attributes has been established to characterize sound but it is difficult to include them in a listening test. A previous study aimed at clustering attributes in families. Thus 3 families of attributes were highlighted, "timbre", "space" and "defects". The overall quality and these three families were evaluated in the listening test presented in this article. The test protocol was based on the Mushra recommendation. However it included three anchors specific to each attribute and no reference in order to evaluate quality instead of fidelity. The aim of the experiment described in this paper was to verify the influence of those 3 attributes on the overall quality in a 5.1 reproduction system. It results that the defects attribute has more influence on the overall quality than the timbre and the space. Moreover the presentation of the three attributes on a same screen adds no bias.

1 Introduction

Before being broadcasted on services, the quality of audio contents has to be evaluated. But, current methods of quality assessment reveal some lacks. Despite the development of spatial technologies, standards do not take into account specific features of spatial sound. The basic audio quality (BAQ) is often the only evaluated attribute. According to ITU-R BS.1534 [1], BAQ is the "global attribute used to judge any and all detected differences between the reference and the object". It would be interesting to obtain some clues on impairments influencing the overall quality. Some attributes, such as coloration, brightness, distortion, localization... have been highlighted by different elicitation methods. However their definitions and their understandings remain a major problem and it is difficult to include them in a listening test [2]. Rather than submitting a list of attributes to the listener, it is possible to gather them in different main sound families. The bias created by specific attributes meanings is therefore reduced. Hence a previous study highlighted 3 sound families for qualifying audio contents: "timbre", "space" and "defects" [3]. Others categories of attributes were defined by studies as timbral, frontal and surround fidelity attributes. These tests showed that timbral fidelity was more correlated to the BAQ than spatial fidelity [4]. For each excerpt, the aim of those experiments was to compare various items to their reference for each of the 4 fidelity parameters. The term fidelity was employed because tests included an explicit reference. One of the requirements for the method tested in this paper was that there were no reference. Nonetheless, the original version, was considered as a hidden reference. The aim of the experiment described in this article was to test a quality evaluation method and to prove the influence, precisely the weight of those attributes families on the overall quality in the context of spatial audio.

2 Attributes families

A previous experiment was run in order to highlight families of sound attributes to evaluate the quality of spatial audio [3]. Tests consisted in presenting a list of attributes (28) and asking assessors to classify them in some categories. No sound was presented in order to create groups independently of audio restitution systems. Two methods were employed: a multidimensional scaling (MDS) and on the other hand a free categorization and a clusters analysis. Both tests obtained the same results and thus three families were defined.

- Defects: are interfering elements or nuisances present in a sound, e.g. noise, distortion, background noise, hum, hiss, disruption

- Space: refers to spatial impression-related characteristics, e.g. depth, width, localization, spatial distribution, reverberation, spatialization, distance, envelopment, immersion

- Timbre: this family is split into 2 subfamilies :
  The first one deals with the sound color, e.g. brightness, tone color, coloration, clarity, hardness, equalization, richness
  The second one composed of homogeneity, stability, sharpness, realism, fidelity and dynamics describes the timbre but can also be related to other characteristics of sound.

3 Listening test

In this study, the 3 attributes, "timbre, space and defects" were included in the listening test.

3.1 Listening conditions

The listening room respected conditions of the recommendation ITU-R BS.1116 [5]. The audio system was a 5.1 restitution system. The five loudspeakers were placed according to the ITU-R BS.775 [6].

3.2 Programme material

Six audio sequences were randomly presented to the assessors. Excerpts were chosen throught film, environment and music to cover a large range of contents. The six sequences were soccer comments, waves and sea sound, movie scene (a fight), music (orchestra, jazz and a turning sound). Each sequence was no longer than twenty seconds according to the recommendation ITU BS.1534 [1]. For each excerpt, six various versions were presented including the original (unprocessed signal), two codecs and three anchors specific to each attributes family. The 6 versions are described in table 1. The spatial anchor was specially defined for this test and was based on anchors used in the litterature [7],[8]. This spatial anchor consisted in a crosstalk between the front right and the surround left channel and the widening of each channel.

"3.5" item was defined as a timbral anchor, "SA" a spatial anchor and "noise" a defects anchor.

Table 1: Description of items.

N°	Abbrevation	Item
1	cod 1	Codec 1
2	3.5	Low pass filtered at 3.5 kHz
3	cod 2	Codec 2
4	noise	Pink noise added
5	o	Original (unprocessed signal)
6	SA	Spatial degradation

3.3 Panel composition

Twenty four “experts” assessors participated in quality tests. They are able to detect impairments in audio signals and they have solid musical background due to their job in audio or musical field. The first test session was made by all the assessor population. However the second part of the test split the panel in two groups.

3.4 Test protocol

The test was decomposed in two sessions. The first one was the evaluation of the overall quality and the second one was the assessment based on the 3 main attributes (“timbre”, “space” and “defects”). The test protocol was inspired by Mushra method (ITU-R BS.1534) [1]. Stimuli were presented simultaneously and assessors scored all items on a quality scale. This test included no explicite reference, thought the original version could be considered as an hidden reference. It was noticed that some biases encountered in standards come from the scale [9] thus the proposed grading scale was without labels except on the end point called “low quality” and “high quality”. No number appeared during the grading, assessors had to place the cursor along the slider. One instruction was given: the stimulus perceived as the best quality had to be scored at the top of the scale. The interface enabled to zoom on the excerpt for listening smaller part of the whole audio stimulus. First, all the listeners evaluated the overall quality (OQ). Then, eleven of them assessed the 3 attributes (timbre, space and defects) in a same time (see Figure 1) whereas the other group evaluated each attribute one after the other in three succesive subsessions (see Figure 2). The aim was to verify the kind of presentation to employ during a listening test. Is the grading affected by the evaluation of the 3 sound families in a same screen? Are the assessors unabled to focus their attention on different attributes as suggested in other studies? [10]

4 Results

The researched interest is focused on the method to evaluate audio quality rather than the score and the ranking of the sequences and processes.

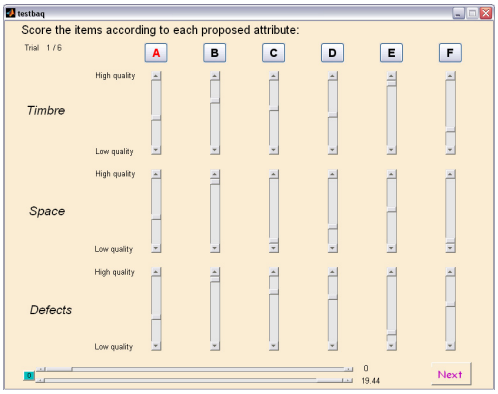

Figure 1: Interface for three attributes presentation

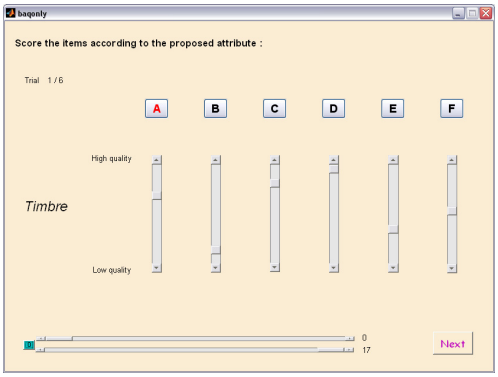

Figure 2: Interface for one attribute presentation

4.1 Attributes presentation

The first thing to notice was the total duration taken by the assessors to complete the test. The single attribute presentation lasted on average 73 minutes whereas the other session took 53 minutes.

Results of the two groups of assessors were compared. A Student test was used to verify the similarity between the scores of both groups. Thus the method of attributes presentation was considered as statistically equivalent. Results obtained by both methods could be merged for the following analyses. Figure 3 shows the similarity between the two methods (method 1: one attribute, method 2: 3 attributes presentation) for the scoring of timbre attribute (mean scores and error bars show 95% confidence interval).

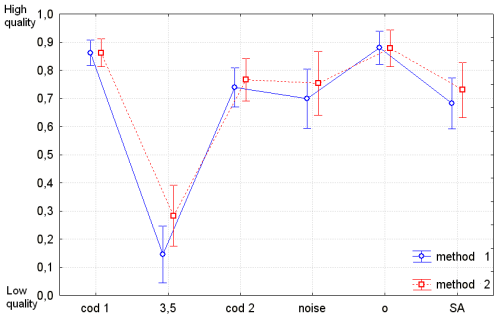

Figure 3: Mean scores and 95% CI of the timbre evaluation for each presentation method

4.2 ANOVA

An ANOVA on each attribute was conducted to highlight factors influencing the scoring. This statistical technique confirmed that the method of attributes presentation had no impact on grading.

Significant effects were revealed by degradations ($p<0.0001$) for each attribute overall quality (OQ), timbre, space and defects). Sequences presented significant effects only for timbre ($F=2.6$, $p=0.032$) and space ($F=7.76$, $p<0.0001$) attributes. Post hoc Tukey's HSD (Honestly Significative Difference) showed that this effect was due to sequences individually and not to a kind of contents (musical and the others excerpts). For exemple, the sequences "sea" and "soccer" were statistically different for spatial attribute whereas they were statistically similar for timbre attribute.

A Tukey's HSD test was performed on degradations for the 4 evaluated attributes. The original version and the "cod 1" had high values and thus were statistically equivalent for all attributes (Tukey values, OQ: 0.96 ; timbre: 0.995 ; space: 0.998 ; defects: 1). By contrast, the rating of the timbral anchor "3.5" is significantly different from the other items for each attribute analysis. For timbre analysis, scores for items "noise", "cod 2" and "SA" were statistically similar. For space analysis, "cod 2" and "noise" ratings are statistically similar with an HSD value of 0.997 and for defects attribute, "cod 2" and "SA" the value was 0.977. Hence, two groups of items were statistically highlighted. The first one consists in the original and the "cod 1" and the second one is composed of "cod 2", "noise" and "SA" but it is dependent on the attributes. An anchor was statistically different from the others items considering the analysis of its associated attribute.

Figure 4 represents mean scores and 95% confidence interval for each attribute evaluation for each item. For both evaluated codings, the obtained notes for each attribute are very close. For exemple, mean values of "cod 2" for all sequences are OQ: 0.72 , timbre: 0.75 , space: 0.76 , defects: 0.76. Moreover "cod 1" is assessed between 0.8 and 0.9 and "cod 2" at about 0.75 for all attributes. The quality of codings used in this test was too high to be included in a test method based on attributes. The overall quality seemed to be sufficient for the assessment of small impairments. With low or intermediate qualities, listeners would be able to detect differences among attributes.

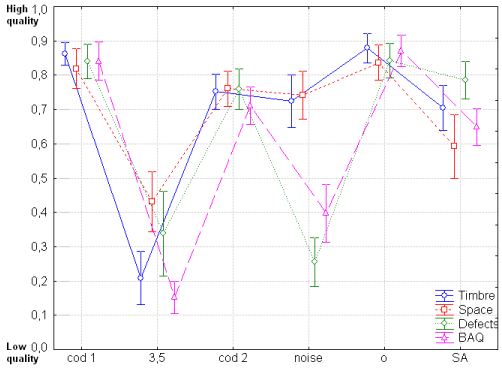

Figure 4: Mean scores and 95% CI of the 4 attributes

4.3 Choice of anchors

Test included 3 anchors, each one focused on an attribute. For the timbre evaluation, the "3.5" item was scored logically in low quality. However this item was also scored in the lower half of the scale for space and defects attributes. The low pass filtered at 3.5 KHz seemed to affect many aspects of sound including space and defects and not only the timbre. The "noise" anchor was the worse item for defects attribute and by contrast, it was scored in high quality for the other attributes (timbre and space). Hence it could be considered as a good anchor for the defects attribute. For the spatial attribute, the spatial anchor (SA) was scored better than timbral anchor ("3.5") and placed in the middle range of the scale, not in low quality (see figure 4). In an other study, spatial anchor was placed in the middle of the quality scale [8]. A question appeared about the possibility to define a spatial anchor scored in low quality. Furthermore it is important to remind that this test was run without explicit reference.

4.4 Correlation between overall quality and attributes

A multiple linear regression was carried out in order to quantify the correlation and the weight of sound attributes with the overall quality.

The results of correlation analysis are presented in the table 2. All variables were correlated with each other. The overall quality was more correlated to the defects (0.90), then timbre (0.87) and space (0.78). Defects attribute was less bonded to the space (0.49) than to the timbre (0.64). Timbre was correlated to space (0.88).

Table 2: Correlation values between overall quality and attributes.

Attributes	Timbre	Space	Defects
Overall quality	0.87	0.78	0.9
Timbre	-	0.88	0.64
Space	-	-	0.49

Results of the regression are summarized in the table 3. The R value (0.985) and the standard error of the estimation indicate that the predicted overall quality and the actual overall quality are very close. The R square value is 0.967 and thus, about 97% of the variance of the overall quality scores can be predicted. Figure 5 representes a scatter plot of the predicted and the observed values of the overall quality. Thus the regression model denotes a high accuracy.

Table 3: Multiple linear regression model summary.

R	R^2	F(3.32)	Std Error of the estimate
0.985	0.97	344.22	0.05

The aim of this study was to find the weight of each attribute on the overall quality score. The values of the standardized regression coefficients (β) are 0,25 for timbre and space attributes and 0.61 for defects which is the attribute that most affectes the overall quality (OQ). The coordinates of the regression equation are given by the unstandardized regression coefficients:

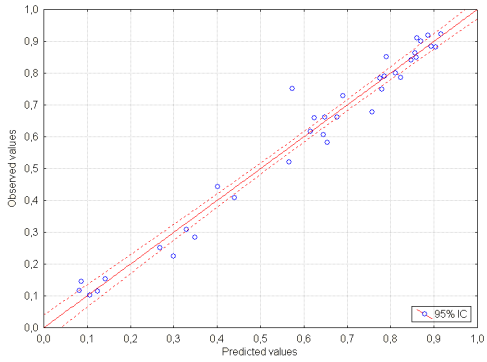

Figure 5: Overall quality, observed vs predicted values

$$OQ = 0,65\ defects + 0,44\ space + 0,3\ timbre - 0.32, \quad (1)$$

These coefficients diverge according to previous studies which concluded that the timbral fidelity has more influence on the basic audio quality than the spatial fidelity [4]. The difference can be explained by the limited number of codings. By consequences, anchors affects strongly the results. Furthermore, the quality was evaluated instead of fidelity (test with no reference). A third attribute called "defects" was introduced and is assessed as the most influential attribute on the overall quality.

5 Conclusion

The listening test method proposed in this paper, was based on the quality evaluation of three sound families, named "timbre, space and defects". Two attributes presentations were tested by assessors, the evaluation of the three attributes simultaneously in one session or the evaluation of attribute in three subsessions sucessively. Results showed that the kind of attributes presentation was not significant. But the three attributes presentation had the advantage of a shorter duration to complete the test. The method included one anchor by attribute. This allowed to verify the well understanding of the attributes definitions by the assessors. The anchors had to be scored in low quality. As mushra test, this method seems to be dedicated to audio with intermediate quality. Impairments on each attribute had to be detected by listeners in order to scores reveal information. The number of evaluated codings was limited. More codecs should be included in the test in order to provide more conclusions. The regression model proposed was accurate. A regression equation was defined and the overall quality could be predicted. This demonstrated the influence of the defects rather than space and timbre on the overall quality. Taking into account those results, a spatial anchor has to be defined and codecs with intermediate quality will be evaluated. Moreover, in the same way, the method is used on headphones with binaural materials.

References

[1] ITU-R Recommendation BS.1534, "Method for the subjective assessment of intermediate quality level of coding systems," International Telecommunications Union, Radio-communication Assembly, Tech. Rep., 2003.

[2] S. Le Bagousse, M. Paquier and C. Colomes, "State of the art on subjective assessment of spatial sound quality," *AES Int. Conf. on Sound Quality Evaluation*, 2010.

[3] S. Le Bagousse, M. Paquier, and C. Colomes, "Families of sound attributes for assessment of spatial audio," *AES 129th Convention*, 2010.

[4] P. Marin, F. Rumsey, and S. Zielinski, "Unraveling the relationship between basic audio quality and fidelity attributes in low bit-rate multi-channel audio codecs," *AES 124th Convention*, 2008.

[5] ITU-R Recommendation BS.1116, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," International Telecommunications Union, Radio-communication Assembly, Tech. Rep., 1997.

[6] ITU-R Recommendation BS.775, "Multichannel stereophonic sound system with and without accompanying picture," International Telecommunications Union, Radio-communication Assembly, Tech. Rep., 1994.

[7] EBU-TECH 3324, "EBU evaluations of multichannel audio codecs," European Broascasting Union, Tech. Rep., 2007.

[8] A. Mason, D. Marston, F. Kozamernik, and G. Stoll, "EBU tests of multi-channel audio codecs," *AES 122th Convention*, 2007.

[9] S. Zielinski, P. Brooks, and F. Rumsey, "On the use of graphic scales in modern listening tests," *AES 123th Convention*, 2007.

[10] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech, "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 968–976, 2005.