

COVER PAGE

Document downloaded by @DAEL

Fri Nov 27 11:15:51 2026
For personal use

When automatic English translation is provided, only the original document is authentic.
The EAA cannot be held responsible of any translation error

Bibliographical reference

Determination of a Relevant Spatial Anchor for Audio Quality Evaluation of Codecs, S. Le Bagousse, M. Paquier, S. Moulin and C. Colomes, Acta Acustica **vol. 102** (Number 2), 2016, pp. 383-388

DOI
https://doi.org/10.3813/AAA.918954

Acta Acustica united with Acustica
Vol. 102 (2016) 383 – 388

Technical and Applied Papers
DOI 10.3813/AAA.918954

Determination of a Relevant Spatial Anchor for Audio Quality Evaluation of Codecs

S. Le Bagousse [1], M. Paquier [1], S. Moulin [2], C. Colomes [2]
[1] University of Brest, Lab-STICC CNRS UMR 6285, 6 avenue Le Gorgeu - 29238 Brest, France.
lebagousse.sarah@free.fr, Mathieu.Paquier@univ-brest.fr
[2] Orange Labs. 4 rue du clos courtel, 35510 Cesson Sevigne, France

Summary
In most tests on audio quality, a "low quality" anchor is provided concurrently with the sequences to be evaluated. A classic generic anchor consists of a low-pass filtered version of the audio reference; but evaluation of specific sound properties (for example spatial features) could require a specific anchor. In a first experiment, we evaluated five spatial anchors to find which was the most appropriate for a test including evaluation of *Space* quality. One of the tested anchors obtained low scores for all the excerpts used. In a second experiment, a copy of this anchor was integrated into an evaluation test of audio codecs, with assessment of the categories *Timbre, Defects*, and *Space*. The scores obtained in this test indicated that this anchor fulfilled the required criteria (a low score for the anchor over all excerpts, the correct range of codec scores for *Space*, and no serious degradation for *Timbre* and *Defec ts*).

PACS no. 43.75.Cd, 43.66.Pn, 43.66.Yw, 43.66.Lj

1. Introduction

Current methods of sound quality assessment have some deficiencies and biases [1, 2]. In particular, recommendations do not take into account specific spatial features of sound, with basic audio quality (BAQ) often being the sole attribute evaluated [3].

A previous study [4] highlighted three categories of sound attributes that could be used to describe audio quality, particularly in the context of an audio codec assessment, these are:

- *Defects*: there 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*: deals with the sound color, e.g. brightness, tone color, coloration, clarity, hardness, equalization, richness .

For sound evaluation tests (e.g. evaluation of audio codecs), these three properties (or categories of sound attributes) could therefore be evaluated in addition to the classic BAQ.

The inclusion of these three properties in an evaluation test, means that three "low quality" anchors (specific to each property) should also be incorporated into the test method. The logical choice for the Timbre anchor could be the common low-pass filtering method [3]. Even if Defects anchor is not common, it can be thought that the addition of noise and/or clicks would be a good choice. Finding an anchor for spatial attributes is more difficult because the rare tested examples of such spatial anchors have been scored in the upper part of the grading scale [5, 6, 7, 8].

A previous study examined the comparative effects of controlled multichannel audio bandwidth limitation and selected down-mix algorithms on audio quality [8]. A hidden reference and two hidden anchors were used in this experiment: a low-pass filtered (3.5 kHz) version and a down-mixed (mono) version of the original recording. The audio quality of the mono version was on average graded as "Fair" (scores between 40% and 60%). On the contrary, the low-pass filtered versions gave scores inferior to 20%). The authors concluded that listeners tolerated degradation of spatial quality due to a limited number of channels better than degradation of timbral quality caused by low-passed filtering. A possible explanation is that a monophonic sequence could be perceived as an absence of spatialization rather than poor spatialization.

In [7], several multichannel audio codecs were evaluated using a Mushra method [3]. A spatial anchor was added to the protocol, consisting of a down-mixed version of the original multichannel content. The spatial anchors always obtained a score superior to 50% when assessed.

In [9], the perceptual importance of the three attributes *Timbral fidelity, Frontal spatial fidelity*, and *Surround spatial fidelity* and their contribution to the basic audio quality

Received 19 June 2015,
accepted 04 December 2015.

© S. Hirzel Verlag · EAA

of low bit-rate surround sound codecs was investigated. The authors used several audio codecs, several excerpts, a classic low-pass filtered version (timbral anchor), and a specific spatial anchor consisting of a reduction of the spatial image width [7]. The scores of the spatial anchor were above 60

In [5], the three categories *Timbre, Defects* and *Space* were included in a test on the assessment of spatial audio quality with binaural contents. Subjects had to listen to four audio codecs, the original version, and the anchors, without an explicit reference. The spatial anchor used (L/R inversion alternated with unaltered stereo and mono reduction) obtained a score of 67%. The score of the spatial anchor is particularly high when no explicit reference is presented: a modified auditory space is often not perceived as degraded. As a consequence, we need to ask how one should define a spatial anchor.

The present study aims at defining an appropriate spatial anchor that could be included in a method of quality assessment dedicated to spatial audio. In the first experiment, several spatial anchors were compared. In the second experiment, a copy of the "best" anchor from the former experiment was integrated into a Mushra-like test in which subjects had to assess quality of codecs on the three categories *Timbre, Defects* and *Space*.

2. Experiment A

2.1. Spatial anchors

An appropriate spatial anchor must fulfil certain criteria:
- It should be evaluated as low quality for the *Space* attribute.
- Similar means should be obtained whatever the type of sound excerpt (e.g. rock, jazz, classical music, movie soundtrack, etc.).
- It should impair mostly the *Space* attribute (rather than *Timbre* and *Defects*).
- It should obtain lower scores than the regular sequences in the tests (for example codecs), but these should lie within the same range overall. An anchor that is too impaired would be assessed so low that all the other sequences would be very well scored in comparison, defeating the purpose of the test.
- It should provide a degraded signal such that the rating results of the Mushra test can be interpreted by people that cannot listen to the test items. It should be easy for everyone to understand the degradation it causes.

In this study we compared several anchors. Some had already been used in previous studies, and some others were developed specifically for this study. The initial audio material and the derived anchor were in 2-channel format. L' and R' represent the processed left and right channels of the anchor, respectively, obtained from the original L and R channels. The 2-channel format was chosen (instead of 5.1, for example) because stereo is still the most frequently used format, and the spatial features of sound are already largely perceptible with only two channels.

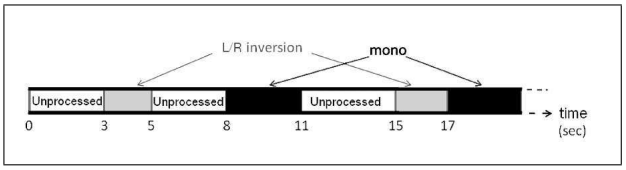

Figure 1. Anchor with dynamic switch of channel.

Anchor A: Mono reduction. This anchor was used in [8], and obtained a score in the middle part of the grading scale.

$$L' = R' = \frac{L+R}{2},$$

with :
L: the initial left channel of the excerpt,
R: the initial right channel of the excerpt,
L': the left channel of the anchor,
R': the right channel of the anchor.

Anchor B: this anchor was a reduction of the sound image (similar to the anchors used in [7] and [9] for multichannel contents). This reduction of the sound image consisted here in adding 25% of the left channel to the right channel and adding 25% of the right channel to the left channel.

$$L' = (1-\alpha)L + \alpha R \quad \text{and} \quad R' = (1-\alpha)R + \alpha L.$$

Here, with $\alpha = 0.25$, L' contains 75% of L and 25% of R, and R' contains 75% of R and 25% of L.

Anchor C: for this anchor used in [5], three periods were alternated: unaltered stereo, L/R inversion, and mono reduction (Figure 1).

Anchor D: for this anchor, a swinging movement was created between the right and left channels. The monophonic signal, M, swung between the right and the left sides following a sine function.

$$L' = L/2 + \alpha M \quad \text{and} \quad R' = R/2 + \beta M,$$

with $M = (L+R)/2$ (monophonic reduction), $\alpha = \sin(2\pi At)^2$ and $\beta = \sin(2\pi At + \pi/2)^2$, with A (spatial modulation frequency) = 0.1 Hz.

These formulae and parameter values were chosen perceptually based on an informal pre-test in order to optimize the spatial inconsistency. It is worth noting that this processing provided a time-dependent level fluctuation.

Anchor E: this anchor was built from a commercial plug-in (with blind confidential processing) chosen according to informal tests in which listeners had reported an impression of rotation that was quite similar (but more drastic) to perceptual spatial degradations sometimes heard with audio codecs.

The anchors C, D and E were chosen because audio codecs can sometimes give an impression of instability or movement of the auditory space.

It is worth noting that the anchor E would not be suitable for a Mushra test, as the signal processing –i.e. the degradation it causes– is not precisely known. However it

Table I. Excerpts used in pre-test (S_{0a} and S_{0b}) and test (S_{1to8}).

S_{0a} (pre-test - Dion)	Pop music with vocals	Stereo
S_{0b} (pre-test - Stairs)	Two people walking down stairs	binaural recording
S_1 (Alanis)	Pop music with vocals	Stereo
S_2 (Barber)	A man having a haircut in a barbershop	binaural recording
S_3 (Bombarde)	A Breton oboe player walking and playing	binaural recording
S_4 (Marimba)	Piece of music played on a marimba	binauralized 5.1 mix
S_5 (Milanof)	Brass band	binauralized 5.1 mix
S_6 (Money)	Instrumental rock music	stereo
S_7 (Star Wars)	Film soundtrack: space vessels and gun shots	binauralized 5.1 mix
S_8 (Tango)	Piece of tango music (bandoneon and strings)	binauralized 5.1 mix

was used in this first experiment to get an impression of the degradation this type of anchor could provide in comparison to the other tested anchors.

In principle, none of these spatial anchors should degrade the attribute categories *Timbre* (color, equalization, etc.) or *Defects* (presence of noises, clicks, etc.). This has been verified in the second experiment of this study.

2.2. Listening test

2.2.1. Stimuli

Ten 20 s-excerpts were used. Three of these were real binaural recordings (with two microphones placed in the ears of a human or a dummy head), four were 5.1 sequences binauralized with a synthesis application, and three were stereo musical recordings. Binaural excerpts were included because this format allows restitution on 360° of azimuth, and even allows for a presentation at different elevations. So the evaluation of spatial features of such sounds is of particular interest.

These excerpts covered a large range of content, including music, film and environmental sounds (Table I), in order to observe any dependence on the type of sound content [8].

The different versions were subjectively matched in loudness by the three experimenters so that the assessments of the listeners were not affected by differences in loudness [10, 11].

For each excerpt, eight versions were presented to the listeners, including the original (unprocessed) signal, the five derived spatial anchors, and two versions processed by compression codecs (table II). These codecs were integrated into the test in order to verify that the anchors were not too close or too far from realist sounds likely to be present in sound quality tests.

2.2.2. Test design

The test consisted in a multi-comparison protocol based on Mushra [3]: all versions of a given excerpt were presented simultaneously on one interface (Figure 2). A continuous grading scale without intermediate labels was proposed for each of the eight versions. The endpoints were denoted low and high quality. No explicit references were presented. Listeners were asked to assess the sound quality exclusively according to the Space attribute ("refers to spatial impression-related characteristics, e.g. depth, width, localization, spatial distribution, reverberation, spatialization, distance, envelopment, immersion"). They had to score the version they considered as the one with best quality out of the eight versions proposed at the maximum of the scale. By clicking buttons, labelled A to H, the assessor could choose the version they would hear. The interface allowed for multiple presentations of the sounds.

When a listener was sure of his assessments, he had to click the "next" button, and the eight next sliders (concerning the versions of the next excerpt) would then be displayed. The versions corresponding to buttons A to H were chosen randomly for each excerpt.

The test began by a 5 minutes pre-test (with two sequences S_{0a} and S_{0b} that were used exclusively in this

Table II. Description of the eight versions of each excerpt proposed to listeners in experiment A.

Original	Unprocessed version
Anchor A	Mono reduction
Anchor B	Reduction of sound image
Anchor C	L/R dynamic switch
Anchor D	Swinging movement I
Anchor E	Swinging movement II
Codec 1	HEAAC (24 kbits/s)
Codec 2	AAC (32 kbits/s)

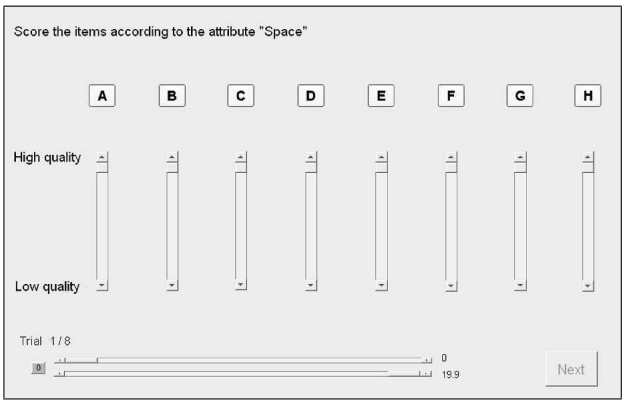

Figure 2. Test Interface. Listeners could listen to the different versions by clicking one of the buttons (A to H), and had to give their assessments by moving the sliders with respect to each version. The two horizontal sliders allowed listeners to set respectively the beginning and the end of the sequence.

pre-test) in order to familiarize the listeners with the spatial impairments. The eight other excerpts were used in the main test.

The test took place in a soundproof booth and sequences were reproduced (with levels ranging between 75 and 80 dB SPL) on a STAX Signature headphone with an SRM-006t amplifier. Listeners were asked not to move the headphones during the entire test period [12]. The whole test lasted 45 minutes. No repeats were conducted.

Twenty-one listeners participated in this test. They were considered as experts according to [3], as they had expertise in detecting impairments, did not report any hearing loss, and had solid audio or musical backgrounds.

2.2.3. Results

After having checked the normality, homoscedasticity and sphericity assumptions, the assessments were analyzed using repeated measures analysis of variance with version (8 levels) and excerpt (8 levels) as within-subjects factors. As illustrated in Figure 3, this analysis showed that the version effect was statistically significant ($F(7,133) = 35{,}4$; $p < 0.0001$). As expected, the "Original" unprocessed version obtained the best score, with a mean equal to 89%. Anchor B (reduction of the sound image) obtained a score equal to 80%; it was therefore assessed as high quality and could not serve our purpose as a low anchor. According to Tukey post-hoc tests, the score of anchor C (dynamic switch), was not statistically different from the one of the two codecs. It should be noted that these codecs "AAC-32" and "HEAACv2-24" are of intermediate quality, and do not degrade the sound space greatly. So anchor C (mean = 60%) would not make a suitable low space anchor. Anchors D (sinus swinging movement) and A (mono reduction) had scores equal to 48% and 49%, respectively, without a significant difference according to a Tukey post-hoc test, either between them or with the codec 1 score. Finally, Anchor E had a score of 34%, which was significantly different from all the other versions ($p = 0.032$ for the smallest difference, which was with Anchor A, according to the Tukey post-hoc test).

Anchor E obtained the lowest scores overall, and the analysis of scores across the different excerpts did not reveal any tendency about a particular efficiency of an anchor for a given type of content. However, the dispersion of scores across excerpts (not visible on figure 3, where the dispersion integrates both subjects and excerpts) was particularly low for anchor E, for which the excerpt with the highest score was only 48% (compared with 62% for anchor D and 67% for anchor A). As a consequence, the type of degradation provided by the anchor E seemed to be the most suitable for all contents.

3. Experiment B

In this experiment a spatial anchor was included in a test in order to verify that it did not degrade the categories *Timbre* and *Defects*.

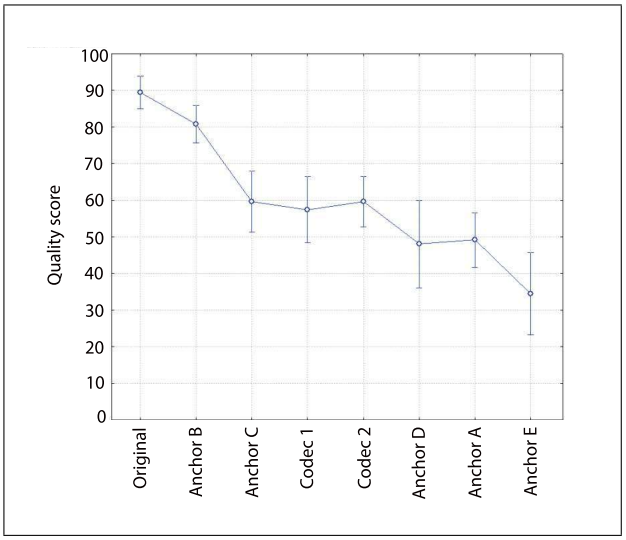

Figure 3. Mean scores for the eight versions, within their 95% confidence intervals (integrating dispersion across excerpts and subjects).

3.1. Spatial anchor

As spatial anchor E obtained the lowest assessment in experiment A and as its precise processing is not available, we developed an anchor as close as possible to this one. The perceptual proximity of the two anchors was confirmed by a discrimination test with 20 listeners, using a 3-Interval 3-Alternative Forced Choice (3I3AFC) response paradigm. During a trial, three intervals were successively presented: one containing an excerpt impaired with an anchor processing and two containing the same excerpt impaired with the other anchor processing.

The original anchor processing E and the copy were randomly assigned to the intervals. The stimulus that was presented once (referred to as the oddball stimulus) could then appear in first, second or third position. After having listened to the 3 stimuli (no repetition was allowed), the subject had to indicate which one of the 3 intervals was the oddball one (the "odd-one-out"). Ten excerpts were used in this test. They were 7s-long parts for each of the ten excerpts used in the experiment A (using the entire 20s excerpts would have make the differentiation task too difficult because of the short auditory memory). Listeners obtained an average score of 32%, which was not significantly different from the chance percentage (33.3%) according to a T-test. This means that listeners were not able to discriminate between the original anchor E and the copy.

For this new anchor, the signals were processed as

$$L' = \frac{L - R}{V} + \alpha M \quad \text{and} \quad R' = \frac{R - L}{V} + \beta M,$$

with $M = (L + R)/2$ and $V = 7$.

For one cycle of $\alpha(t)$ and $\beta(t)$,

$\alpha(t) = 1 - \delta t$ if $t \in [0, p]$ seconds,

$\alpha(t) = 0$ if $t \in [p, p + n]$ seconds,

$\alpha(t) = \delta.t$ if $t \in [p + n, 2p + n]$ seconds,

Table III. Description of the eight versions of each excerpt proposed to listeners in experiment B.

Original	Unprocessed version
Anchor T	3.5 kHz low-pass filtering
Anchor D	addition of pink noise and clicks
Anchor S	Swinging movement (trapezium)
Codec C1	HE-AACv2 (40 kbits/s)
Codec C2	AMR WB+ (48 kbits/s)
Codec C3	MP3 (64 kbits/s)
Codec C4	AAC (32 kbits/s)

$\alpha(t) = 1$ if $t \in [2p + n, 2(p + n)]$ seconds,
$\beta(t) = 1 - \alpha(t)$,
$\delta = 1.7 \cdot 10^{-5}$; $n = 1.93$; $p = 1.34$.

Hence, the duration of one cycle for $\alpha(t)$ and $\beta(t)$ was 6.54 s (see Figure 4). The modulations $\alpha(t)$ and $\beta(t)$ were applied on the whole excerpts. It is worth noting that a trapezoid modulation is used for this anchor (as the modulation of the anchor E used in experiment A seemed to be itself trapezoid).

3.2. Listening test

In order to verify that the spatial anchor did not impair non-spatial properties of sounds, we designed a listening test in which listeners had to assess stimuli across the three categories *Timbre, Defects* and *Space*. The stimuli consisted of eight versions of ten excerpts. The eight versions are listed in Table III. In order to evaluate our spatial anchor in a more realistic test (with realistic impairments), we also used four audio codecs. Moreover, as listeners were asked about the three categories *Timbre, Defects* and *Space*, the anchors specific to *Timbre* and *Defects* were also included in the test [6].

The test interface for experiment B was similar to the one for experiment A, but there were three rows of sliders (one row for each category) as recommended in [13]. The excerpts were identical to those used in experiment A. Twenty expert listeners participated in the test.

3.3. Results

Figure 5 indicates the scores for the three categories *Timbre, Defects*, and *Space*, across the eight proposed versions.

For each of the three attributes, the original version obtained the highest score (> 83%). Each anchor obtained the lowest score when the assessment concerned the category associated with it. In particular, anchor S, which is the object of this study, obtained the score 15% for the assessment of the *Space* attribute. This score is very low, which makes this a 'good' anchor. Otherwise, (i) the range of answers concerning the *Space* category was as expected (the codec assessments are not pushed on the score of the original version); (ii) anchor S had low scores for the assessment of category *Space*, but did not have low scores for the *Timbre* and *Defects* categories (the scores obtained by Anchor S for *Timbre* and *Defects* are in the range of those

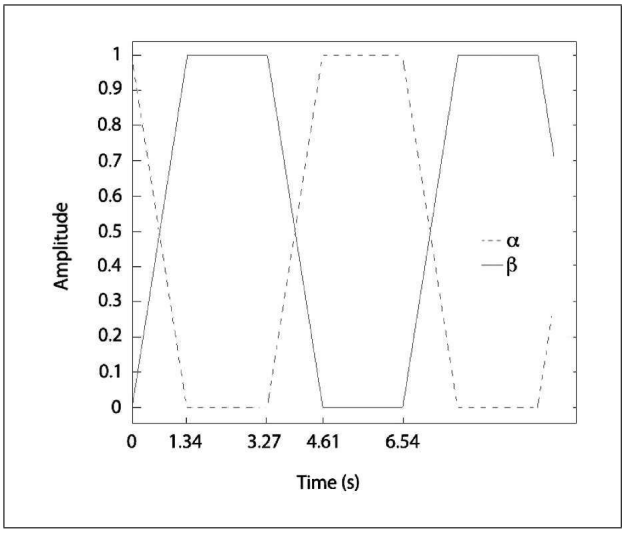

Figure 4. Slopes $\alpha(t)$ and $\beta(t)$ used in the spatial anchor of experiment B.

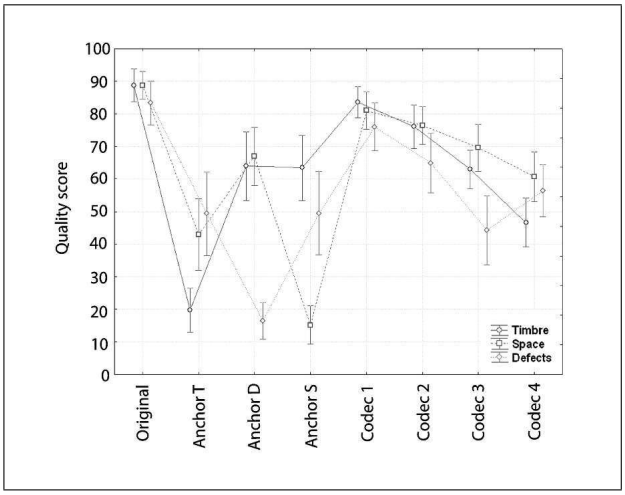

Figure 5. Mean scores for the eight versions, within their 95% confidence intervals (integrating dispersion across excerpts and subjects).

obtained by the codecs); (iii) the highest score of the spatial anchor was only 28% (for the excerpt S2: the barber), showing that this anchor is efficient whatever the excerpt. As a consequence, anchor S developed in this paper fulfills the required criteria to be included in a test method for quality assessment, at least in the context of codec evaluation.

4. Conclusion

In the first experiment, five spatial anchors were evaluated and compared in order to determine a type of anchor that would give low scores for the *Space* category assessment, and a low influence of the excerpt. One of the tested anchors fulfilled these requirements. In the second experiment, it was verified that this type of spatial anchor did not impair non-spatial properties of sounds. To this end, a listening test was used in which listeners had to assess stimuli across the three categories *Timbre, Defects*, and

Space. The stimuli consisted of eight versions (original, four audio codecs and one anchor for each category) of ten excerpts. Whatever the excerpt, the score obtained by the spatial anchor was very low, exclusively for the assessment of the *Space* category. So the spatial anchor developed in this study (based on a trapezium swinging movement) fulfilled all the required criteria for a low-quality spatial anchor.

References

[1] S. Zielinski: On some biases encountered in modern listening tests. Spatial audio & sensory evaluation techniques conference, Guilford Rovaniemi, UK, 2006.

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

[3] ITU-R BS.1534: Method for the subjective assessment of intermediate quality level of coding systems. Technical Report International Telecommunications Union, Radio-communication Assembly, 2003.

[4] S. Le Bagousse, M. Paquier, C. Colomes: Categorization of sound attributes for audio quality assessment - a lexical study. J. Audio Eng. Soc. **62** (2014) 736–747.

[5] S. Le Bagousse, M. Paquier, C. Colomes, S. Moulin: Sound quality evaluation based on attributes - application to binaural contents. AES 131th Convention, New York, USA, October 20-23 2011.

[6] S. Le Bagousse, M. Paquier, C. Colomes: Assessment of spatial audio quality based on sound attributes. presented at Acoustics 2012, Nantes, 2012.

[7] A. Mason, D. Marston, F. Kozamernik, G. Stoll: Ebu tests of multi-channel audio codecs. presented at the 122th AES Convention, Vienna, May 5-8, 2007.

[8] S. Zielinski, F. Rumsey, S. Bech: Comparison of quality degradation effects caused by limitation of bandwidth and by down-mix algorithms in consumer multichannel audio delivery systems. presented at the 114th AES Convention, Amsterdam, 2003.

[9] P. Marins, F. Rumsey, S. Zielinski: Unravelling the relationship between basic audio quality and fidelity attributes in low bit-rate multi-channel audio codecs. presented at the 124th AES Convention, Amsterdam, May 17-20, 2008.

[10] A. Gabrielsson, U. Rosenberg, H. Sjogren: Judgments and dimension analyses of perceived sound quality of sound-reproducing systems. J. Audio Eng. Soc. **55** (1974) 854–61.

[11] V. Koehl, M. Paquier: A comparative study on different assessment procedures applied to loudspeaker sound quality. Applied Acoustics **74** (2013) 1448–1457.

[12] M. Paquier, V. Koehl: Discriminability of the placement of supra-aural and circumaural headphones. Applied Acoustics **93** (2015) 130–139.

[13] S. Le Bagousse: Elaboration d'une methode de test pour l'evaluation subjective de la qualite des sons spatialises. Dissertation. University of Brest, 2014.