Spatial audio quality in regard to 3D video

Samuel Moulin, Rozenn Nicol, Laetitia Gros

ACOUSTICS 2012

Spatial audio quality in regard to 3D video

S. Moulin, R. Nicol and L. Gros

France Télécom - Orange Labs, FT/OLNC/RD/TECH/OPERA/TPS, 2 Av. Pierre Marzin, 22300 Lannion, France
rozenn.nicol@orange.com

3D movies provide an improved immersion in terms of visual perception. As for the associate audio channels, most of them are mixed for the conventional format of "multichannel 5.1". It should be considered that today there are various ways of listening to 5.1 audio content, either over loudspeaker arrays (for instance ITU standard 5.1 set-up), or over headphones. Recently, sound projectors were introduced, in order to render surround sound with compact equipments. The choice of the solution to render multichannel audio has obviously an impact on the perception of the sound. In addition, this latter will also depend on whether a visual content is presented in combination with the audio content. This paper will re-examine these issues in the new context of 3D video. The perception of the audiovisual scenes (audio, video and cross-modal perception) is assessed by a listening test for a set of audiovisual excerpts of a 3D movie.

1 Introduction

Since a few years, the emergence of 3D multimedia contents represents a major evolution. In this technological race, video catches most of the attention and the question of a suitable audio is poorly investigated. Due to its popularity, 5.1 surround system is the most often used sound spatialization technology. But the 5.1 format is far from being the only solution. There is a wide range of alternative 3D audio technologies like binaural technologies, Wave Field Synthesis or Higher Order Ambisonic for example. Now, it is well-known that all sound spatialization technologies are not equivalent, in terms of rendering of each dimension of the sound space (azimuth, elevation, distance), and particularly in terms of depth rendering. More precisely, WFS is known to render the depth of sound sources [1,2,3].

The question of interaction between the audio and video rendering already arose with 2D video (either for psychological approaches [4], or quality assessment [5]) but it deserves to be re-assessed for stereo display. Indeed, disparities between right and left images provide depth reproduction which is the new element brought by the 3DTV technique.

When putting together 3D video and 3D audio, it is clearly of considerable interest to compare the audiovisual perception as a function of the sound spatialization technology.

As a first step, this paper will focus on spatial sound systems which are commonly used to reproduce 5.1 audio contents. The most straightforward solution is a loudspeaker array, based for instance on the ITU standard 5.1 set-up. An alternative is sound reproduction over headphones with a down-mix processing. Recently, sound projectors were introduced in order to render surround sound with compact equipments. It is intended here to measure the impact of the solution chosen to render multichannel audio, on the perception of 3D audiovisual content. Among these existing systems, is there a more suitable solution? Can the visual perception of stereoscopic images be influenced by a particular sound reproduction system? To answer these questions, subjective test should be performed, but first a proper methodology must be defined.

Indeed, most of the recommended 2D subjective assessment methodologies address only one single modality (ITU-R BT.500 and BT.1788 for video quality [6,7], or ITU-R BS.1284, BS.1534 and BS.1116 for audio quality [8,9,10]). Some recommendations make suggestions for evaluating one modality in an audiovisual context or in the presence of an accompanying signal in the other modality (ITU-R BS.775-2 for multichannel audio with accompanying picture [11], ITU-R BS.1286 for the testing of audio systems with accompanying image [12], and ITU-T P.910 for evaluating the one-way overall video quality for multimedia applications such as videoconferencing [13]). Only P.911 and P.920 recommendations [14,15] applied to audiovisual subjective assessment, in a non-interactive context or in an interactive one. But, if visual depth is added to the initial 2D video, these latter should be potentially rethought, or at least questioned [16]. For example, Wei Chen suggested taking into account additional assessment attributes such as depth sensation, or visual comfort [17]. Similar issues are encountered with spatial sound: there is a lack of methods for the assessment of spatial audio quality. Indeed, standard methods [9,10] do not take into account specific features of spatial sound and assessments are limited to the overall sound quality. That is why many studies are focused on the development of new methodologies for quality assessment of sound spatialization [18,19,20,21]. Anyway, it should be noticed that, up to now, audio and video assessments remain independent: 3D video assessment is studied on one side, and spatial sound on another. We will consider the recent work on both sides to design our experimental test for the subjective assessment of 3D audiovisual contents.

This paper will describe a subjective test in which 15 excerpts of a 3D movie are presented to assessors. The 3D movie is mixed in 5.1 surround audio format. The aim is to present audiovisual sequences with 3 different sound reproduction systems (5.1 surround system, sound projector and headphones). For each sound reproduction system, spectators have to evaluate the fifteen audiovisual sequences through six different criteria focusing either on video, or audio, or the combination of audio and video. First, the experimental protocol is given in part 2. Then, the results are presented and discussed in part 3.

2 Test Design

2.1 Environment

Video excerpts are displayed on a LG 47LW5500 47" stereoscopic LCD screen with a 1920x1080 resolution. Spectators need to wear polarized glasses to see stereoscopic effects throughout the test duration. The passive technology is chosen because it appears more comfortable for the audience in comparison to active stereoscopic technology due to the weight of glasses. Moreover, polarized glasses are brighter than active shutter glasses.

Regarding to sound reproduction, three systems are used for the play-black of 5.1 sounds: Genelec 8040A Bi-amplified 5.1 multichannel system, Yamaha YSP-1 sound

projector and AudioTechnica ATH-AD700 open headphones. All the sound reproduction systems are upstream controlled by a Terratec Phase 26 USB external sound card.

The test is performed in an acoustically treated room at Orange Labs. In this room, the background noise level is less than 30 dBA and the background room illumination is less than 20 lux as recommended in ITU-T P911 [14]. The recommended viewing distance is three times the height of the screen for HDTV [6]. According to the monitor dimensions, the spectator sits at 1.85 meter from the 3DTV. The sound projector is placed under the TV and the multichannel audio system is placed around the listener. The 5 loudspeakers are located in accordance with the ITU BS.775 recommendation [11] and are 1.90 meter distant from the sweet spot.

2.2 Stimuli

Fifteen audiovisual sequences are extracted from a 3D documentary about the boxer Jean-Marc Mormeck. Each excerpt lasts between 11 and 17 sec. The sequences were selected taking into account the maximum acceptable disparity between left and right views [22]. It was also intended to illustrate the various relationships between audio content (speech, dialogue, environmental sounds, background music, etc.) and video content (indoor/outdoor, number of characters, object and camera motion speed, etc.). Sequences characteristics are presented in Table 1.

Table 1: Description of selected sequences.

Seq	Place	Motion	Music	Speech	Sound effects	Max disparity (cm)
1	Outdoor	Dyn	Yes	-	Birds*	1.3
2	Outdoor	Dyn	Yes	Breath	Birds* Footstep	0.5
3	Outdoor	Dyn	Yes	Breath	Birds* Footstep	1.3
4	Indoor	Dyn	Yes	-	Applause * Punch	1.0
5	Indoor	Static	Yes	Yes	-	1.1
6	Indoor	Static	Yes	-	Punch	1.2
7	Indoor	Static	Yes	Voiceover	-	1.3
8	Indoor	Static	Yes	Voiceover	-	0.5
9	Indoor	Static	Yes	Voiceover	-	0.7
10	Indoor	Dyn	Yes	-	-	1.5
11	Indoor	Dyn	-	-	Applause* Punch	1.2
12	Indoor	Dyn	Yes	-	-	1.5
13	Indoor	Static	Yes	Voiceover	-	1.1
14	Indoor	Dyn	Yes	-	Applause *	1.1
15	Indoor	Static	-	Voiceover	-	1.1

Surround sound effects have been added in post-production and are annotated (*) in Table 1.

Video files are encoded using H264 codec with an average bit rate of 30 Mbps (25 frames/s).

All audio files are uncompressed PCM files with an original bit rate of 6912 kbps for 6 channels files and 1536 kbps for down-mixed 2 channels files. In both cases the sampling frequency is 48 kHz.

2.3 Panel composition

The panel consists of 30 participants (21 women and 9 men) whose average age is 32.7 years. Among them, 28 spectators have experience in listening tests, but none of them took part to a subjective test with audiovisual content.

The number of participants is deliberately higher than recommended by the BT-500 (at least 15 observers). Indeed, in the context of 3D video, Chen highlights the need to increase the number of observers because of the instability of viewers' opinion [17].

2.4 Test procedure

The test is divided into three sessions. Each session is dedicated to one system of sound reproduction, *i.e.* either the 5.1 multichannel system, or the headphones, or the sound projector. The order of presentation of sound reproduction systems changes every five spectators.

For each session, audiovisual excerpts are presented three times to spectators so they can focus successively on video, audio and audiovisual properties and assess different criteria:

- During the first presentation, spectators have to assess the following video characteristics: **degree of visual depth** and **viewing comfort**
- During the second presentation, participants have to assess the following audio properties: **degree of sound spatialization** and **listening comfort**.
- Finally, at the third presentation, the audiovisual scene should be assessed as a whole. Criteria are: **degree of coherence** between sound and image and the **degree of immersion** in the audiovisual scene.

After each presentation, participants rate the two corresponding criteria on 5-point scales of which extremes are labelled: "uncomfortable/comfortable" for comfort relating criteria and "low/high" for others.

Thus, for each excerpt of each session (45 trials in total), the assessor have to do six ratings.

Prior to the test, assessors perform a short training task with 4 excerpts selected from the same 3D documentary. With this training, participants can familiarize themselves with the stimuli and the test procedure. The average total test duration is 92.5 min.

3 Results

A variance analysis (ANOVA) is performed on individual scores (between 0 for uncomfortable/low and 4 for comfortable/high), obtained for all criteria considering two between-group factors: "Sound reproduction system" (three levels) and "Sequences" (fifteen levels). Then correlations between the six criteria are also studied.

3.1 Video criteria

Regarding the **degree of visual depth**, Figure 1 shows mean scores and associated 95 % confidence intervals, obtained for each sound reproduction system and each

sequence. It should be noticed that spectators perceived depth difference between sequences ($F(14,406)=11.27$, $p<0.001$). The sound reproduction system seems to not influence the visual depth perception ($F(2,58)=0.48$, $p=0.62$), whatever the sequence considered ($F(28,812)=0.69$, $p=0.88$).

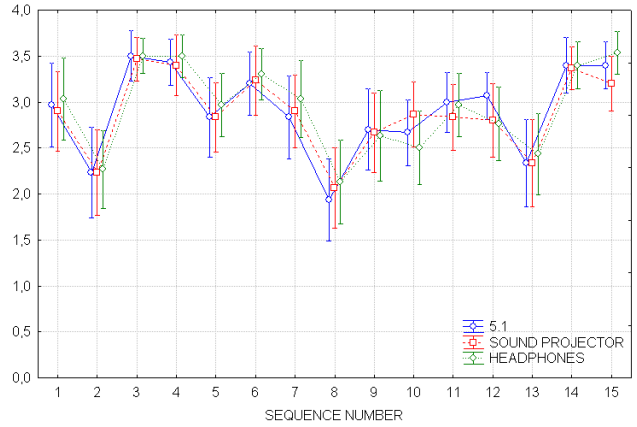

Figure 1: Means and 95% confidence intervals for degree of visual depth criterion.

It is interesting to compare mean scores obtained for perceived visual depth (Figure 1) with the maximum disparities among sequences (Table 1). At first glance, both results seem consistent. Indeed, Figure 1 shows that lower scores are obtained for the 2nd ant the 8th sequence. Nevertheless, the perceived visual depth score is not exclusively due to the maximum disparities. For instance, the maximum disparity of the sequences 13 and 14 is equal (1.1cm) but the perceived visual depth differs. A possible explanation is that even if maximum disparity is strong, participants reported for sequence 13 that they perceived the visual depth as artificial, which can lower their assessments. On the contrary, if visual cues enhance linear perspectives, the perceived visual depth can be upgraded (sequence 4 and 14).

Regarding the **comfort of visualization**, Figure 2 shows mean scores and associated 95 % confidence intervals, obtained for each sound reproduction system and each sequence.

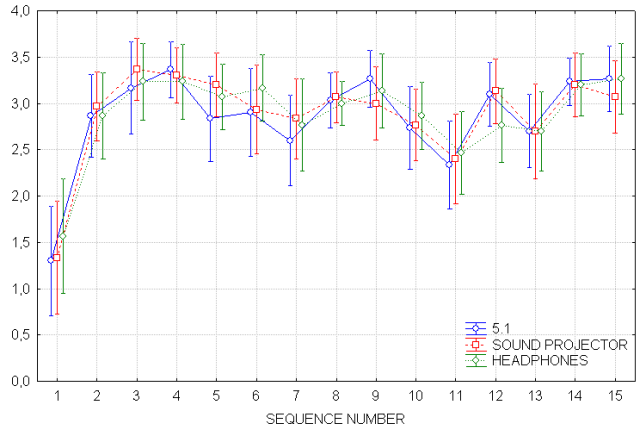

Figure 2: Means and 95% confidence intervals for comfort of visualization criterion.

As for the degree of visual depth, the only significant effect concerns sequences ($F(14,406)=10.21$, $p<0.001$). This effect is mainly due to the first sequence, which is perceived as uncomfortable for all sound reproduction technologies ($F(2,58)=0.24$, $p=0.78$). Assessors judged this particular excerpt as uncomfortable because of fast camera motions.

3.2 Audio criteria

Figure 3 depicts the mean scores concerning the **degree of sound spatialization**. Apart from a weak effect of the sequence on the impression of sound spatialization ($F(14,406)=3.29$, $p<0.001$), it appears that the perception of spatialization essentially depends on the sound reproduction system ($F(2,58)= 9.88$, $p<0.001$). Figure 3 shows that spatialization is generally perceived as more impressive with headphones than with a 5.1 system and that the sound projector is judged lower than the two other technologies, whatever the sequence ($F(28,812)=0.88$, $p=0.64$).

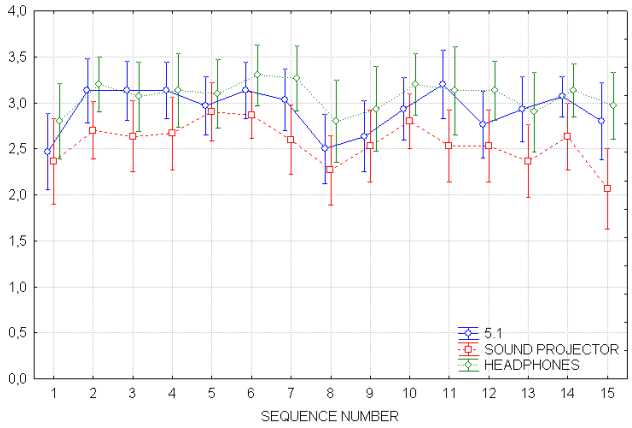

Figure 3: Means and 95% confidence intervals for degree of sound spatialization criterion.

Nevertheless, this strong effect of sound reproduction system is not found for the **listening comfort** criterion ($F(2,58)=2.54$, $p=0.09$). It is observed on Figure 4 that all sound technologies are perceived as quite comfortable: most of mean scores are between 3.0 and 3.5.

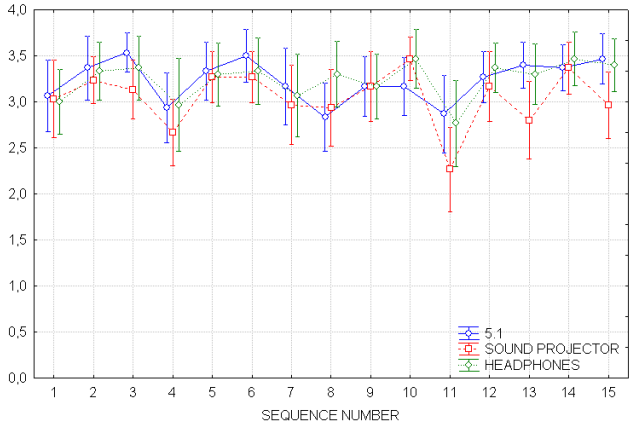

Figure 4: Means and 95% confidence intervals for listening comfort criterion.

Moreover, there is an effect of sequence on listening comfort ($F(14,406)=4.76$, $p<0.001$). Sequences 4 and 11 are perceived as the less comfortable. It can be remarked that these sequences have been post-processed to enhance surround sounds (Table 1).

3.3 Audiovisual criteria

Regarding the **degree of consistency**, the ANOVA shows a little but significant effect of sound reproduction system ($F(2,58)=4.41$, $p<0.05$) as well as a significant effect of sequence ($F(14,406)=4.44$, $p<0.001$). But there is no interaction between these two factors ($F(28, 812)=0.66$, $p=0.91$).

As for the **degree of immersion**, the only significant effect concerns sequences ($F(14,406)=4.28$, $p<0.001$). There is no significant effect of the sound reproduction technology ($F(2,58)=2.69$, $p=0.08$) on the assessment of this attribute.

When these judgments are compared to the previous results (Sections 3.1 and 3.2), it appears that degrees of consistency and immersion are not explained in a trivial way by either visual depth or sound spatialization or visual/listening comfort criteria. Therefore, a correlation analysis is performed and the results are presented in Table 2. V1 and V2 are the video criteria **visual depth** and **comfort of visualization**. A1 and A2 are respectively the **sound spatialization** and **listening comfort** criteria. AV1 is the **consistency** criterion and AV2 is related to **immersion**.

Table 2: Correlation analysis over assessment criteria.

Criterion	V1	V2	A1	A2	AV1	AV2
V1	1	0.41	0.23	0.19	0.36	0.46
V2	-	1	0.20	0.25	0.38	0.46
A1	-	-	1	0.62	0.46	0.43
A2	-	-	-	1	0.57	0.49
AV1	-	-	-	-	1	0.67
AV2	-	-	-	-	-	1

This analysis confirms the lack of clear correlation between criteria. The degree of audiovisual immersion would be more related to the consistency between sound and image (0.67) than to the degree of visual depth or sound spatialization.

4 Conclusion

In this paper, a subjective test is carried out in order to compare the 3D audiovisual perception as a function of the sound spatialization technology. The aim is to measure the impact of three sound reproduction systems chosen to render multichannel audio (5.1 surround system, sound projector and headphones), on the perception of 3D audiovisual content. For each sound reproduction system, spectators have to evaluate audio, video and audiovisual criteria.

Degree of visual depth and comfort of visualization are the criteria for the video part. The possible influence of sound reproduction system on visual depth or on comfort of visualization hasn't be proved in this experiment. But the participants have judged those two criteria as independent. Assessors are able to discriminate different degrees of visual depth between sequences. Their judgments are probably influenced by the maximum disparity between left and right images in each sequence but not exclusively. Indeed, visual depth perception can be attenuated or enhanced by different visual cues like linear perspective for instance. Nevertheless, visual depth perception does not impact the comfort of visualization for selected sequences.

The analysis of audio criteria results shows that, in terms of degree of sound spatialization, headphones listening is rated higher than 5.1, which is itself preferred to the sound bar. Regarding to the listening comfort, all systems are quite equivalent.

Assessment also covers audiovisual criteria (degree of consistency and degree of immersion) but there is no significant effect of sound reproduction systems on these criteria. A correlation analysis shows that all criteria are independent. Immersion seems to be more correlated to audiovisual consistency than to other criteria.

Further work should explore alternatives 3D audio technologies. Wave Field Synthesis is an attractive solution since it potentially impacts the consistency perception by adding audio depth rendering.

Acknowledgments

The authors would like to thank Jérôme Fournier, and Jean-Charles Gicquel, for their technical advices on stereoscopic images. A special thank is given to Mickael Bonin for the development of an audiovisual assessment interface.

References

[1] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by Wave Field Synthesis", *Journal of the Acoustical Society of America*, 93(5), 2764–2778 (1993)

[2] G. Theile, and H. Wittek, "Wave Field Synthesis: A promising spatial audio rendering concept", *Acoustical Science & Technology*, 25(6), 393–399 (2004).

[3] C. Renard, "Analyse objective et subjective d'une technique de rendu sonore 2D sur une zone d'écoute étendue, l'holophonie, en vue de réaliser un mur de téléprésence", *Master Thesis Report*, Université du Maine (2000)

[4] S. Komiyama, "Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems", *Journal of the Audio Engineering Society*, 37(4), 210–214 (1989)

[5] J. G. Beerends, and F. E. De Caluwe, "The influence of video quality on perceived audio quality and vice versa", *Journal of the Audio Engineering Society*, 47(5), 355–362 (1999)

[6] ITU-Recommendation BT.500-12: "Methodology for the subjective assessment of the quality of television

pictures", *International Telecommunications Union, Radio-communication Assembly* (2009)

[7] ITU-Recommendation BT.1788: "Methodology for the subjective assessment of video quality in multimedia applications", *International Telecommunications Union, Radio-communication Assembly* (2007)

[8] ITU-Recommendation BS.1284-1: "General methods for the subjective assessment of sound quality", *International Telecommunications Union, Radio-communication Assembly* (2003)

[9] ITU-Recommendation BS.1534-1: "Method for the subjective assessment of intermediate quality level of coding systems", *International Telecommunications Union, Radio-communication Assembly* (2003)

[10] ITU-Recommendation BS.1116-1: "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems", *International Telecommunications Union, Radio-communication Assembly* (1997)

[11] ITU-Recommendation BS.775-2: "Multichannel stereophonic sound system with and without accompanying picture", *International Telecommunications Union, Radio-communication Assembly* (2006)

[12] ITU-Recommendation BS.1286: "Methods for the subjective assessment of audio systems with accompanying picture", *International Telecommunications Union, Radio-communication Assembly* (1997)

[13] ITU-Recommendation P.910-3: "Subjective video quality assessment methods for multimedia applications", *International Telecommunications Union, Radio-communication Assembly* (2008)

[14] ITU-Recommendation P.911: "Subjective audiovisual quality for multimedia applications", *International Telecommunications Union, Radio-communication Assembly* (1998)

[15] ITU-Recommendation P.920-2: "Interactive test methods for audiovisual communications", *International Telecommunications Union, Radio-communication Assembly* (2000)

[16] S. Pastoor, "Human factors of 3DTV: an overview of current research at Heinrich-Hertz-Institute Berlin", *Stereoscopic Television IEE Colloquium, 11/1-11/4* (1992)

[17] W.Chen, J. Fournier, M. Barkowsky, and P. Le Callet, "New requirements of subjective video quality assessment methodologies for 3DTV", *Fifth International Workshop on Video Processing and Quality Metrics ˘ VPQM 2010*, Scottsdale, Arizona, U.S.A (2010)

[18] J. Berg, and F. Rumsey, "Systematic evaluation of perceived spatial quality", *Proceedings of the 24th AES International Conference on Multichannel Audio* (2003)

[19] J. Berg, "Evaluation of perceived spatial audio quality", *Proceedings of the 9th World Multi-Conference on Systemics, Cybernetics and Informatics*, 4(2), 10–14 (2005).

[20] G. Lorho, "Evaluation of spatial enhancement systems for stereo headphone reproduction by preference and attribute rating", *Presented at* the *118th AES Convention*, Barcelona, Spain (2005)

[21] S. Le Bagousse, C. Colomes, M. Paquier, and S. Moulin, "Sound quality evaluation based on attributes – application to binaural contents", *Presented at* the *131th AES Convention*, New-York, NY, U.S.A. (2011)

[22] S. Yano, M. Emoto, and T. Mitshuhashi, "Two factors in visual fatigue caused by stereoscopic HDTV images", *Display* 25, 141–150 (2004)