Noname manuscript No.
(will be inserted by the editor)

Subjective Quality Assessment of Multichannel Audio Accompanied With Video in Representative Broadcasting Genres

Maximo Cobos · Jose J. Lopez · Juan M. Navarro · German Ramos

Received: date / Accepted: date

Abstract Immersive broadcasting applications have received a lot of attention in the last years. In this context, the development of advanced HDTV and 3DTV formats are being successfully adopted by the consumer market, having a strong impact in the way that traditional broadcasting contents are displayed to final users. Together with the above advances in video technology, multichannel spatial audio has also experienced a considerable impulse within the audiovisual industry. However, the need for specific production tools and loudspeaker set-ups corresponding to multiple competing audio formats seems to be an important factor affecting their adoption by the consumer community. Moreover, it is well-known that the perceived audio quality is highly influenced by the reproduction context, where the existing multimodal interaction between audio and video plays a very important role. This paper presents a formal evaluation of the perceived sound quality provided by several spatial audio formats accompanied with video in the context of television broadcasting. Stereo, advanced surround formats and 3D binaural sound are evaluated considering a set of representative broadcasting contents (sports, movies, music and animation) to assess their impact on the perceptual attributes contemplated within the international recommendations.

Keywords Subjective Quality Assessment · Multichannel Audio · Broadcasting

1 Introduction

The widespread deployment of 3D theaters, new generation multimedia devices and 3DTV is gradually bringing immersive environments closer to final users and consumers. Many research efforts are currently being oriented to the development of new technologies devoted to immersive communication, virtual reality and interactive media [20,36]. As a result, providing high audiovisual quality through the combination of emerging audio and video formats is a major objective in audiovisual research and future broadcasting applications [61,63,37].

The development of immersive multimedia environments is highly linked to spatial audio reproduction [45,51]. Stereo sound systems, considered as the simplest approximation to spatial audio, have been utilized throughout the last 80 years as an added value in sound recordings, specially for music material [28]. Together

M. Cobos
Computer Science Department, Universitat de València
Av. de la Universitat, s/n, 46100 Valencia (Spain).
Tel.: +34-9635-43959
Fax: +34-9635-44768
E-mail: Maximo.Cobos@uv.es

J. J. Lopez
iTEAM Institute, Universitat Politécnica de València
Camino de Vera, s/n, 46025 Valencia (Spain).
Tel.: +34-9638-79714
Fax: +34-9638-77309
E-mail: jjlopez@dcom.upv.es

J. M. Navarro
Advanced Telecommunications Group, San Antonio's Catholic University of Murcia
Campus de Los Jernimos, 135, 30107 Guadalupe (Spain).
Tel.: +34-9638-79714
Fax: +34-9638-77309
E-mail: jmnavarro@ucam.edu

G. Ramos
ITACA Institute, Universitat Politécnica de València
Camino de Vera, s/n, 46025 Valencia (Spain).
Tel.: +34-9638-79714
Fax: +34-9638-77309
E-mail: gramosp@eln.upv.es

with the entertainment industry, stereo sound evolved to surround sound systems, which provide a better spatial sensation than stereo by using more reproduction channels [34]. In fact, the strong link between audio and video has governed the evolution of spatial audio during the last decades, both in theaters and broadcasting applications. A clear example of this connection is the popularity that 5.1 Surround gained with the spread of the DVD and the wide penetration of home-cinema systems [54]. Thus, many audio reproduction techniques and processing methods are continuously appearing to support the advances in the production, coding, transmission and reproduction of audiovisual material [25].

Although the general advantages of using multichannel audio formats in broadcasting seems to be quite clear [56,35], the great variety of audiovisual contents might cause substantial differences in the perceived subjective quality. It has already been shown that different loudspeaker set-ups have a strong influence on TV user experience [55]. However, to the best of the authors' knowledge, no previous works have been focused on the impact that audiovisual content types have on audio perception when conventional and advanced spatial reproduction systems are considered. Although in [64] it was suggested that the presence of video had a small effect on audio quality assessment, only a 5.1 set-up was considered, leaving unclear which multichannel audio formats are preferred according to the displayed content type. In fact, the perceptual attributes governing spatial audio quality might be highly influenced by the contents of the reproduced audiovisual material, thus, it becomes quite difficult to assess the benefits added by certain audio formats within a complete audiovisual context. Despite the fact that several well-known procedures exist [21], the joint assessment of audio and video quality is not a straightforward issue [26]. On the one hand, there is a complex interaction between auditive and visual stimuli in multimodal perception that makes it very difficult to isolate independent quality factors [40]. On the other hand, this complexity can be even higher with certain types of audiovisual contents or the addition of interactive features. Moreover, although several international agencies such as the *International Telecommunication Union* (ITU) [10,4] or the *European Broadcasting Union* (EBU) [13] have addressed the problem of evaluating subjective audio and video quality independently, there are existing contradictions in the required experimental conditions that hinder the evaluation task.

In this paper, we present a formal evaluation of the subjective audio quality provided by several multichannel audio formats accompanied with picture. Diverse types of representative content material in broadcasting (sports, movies, music and animation) are considered to study the effect that they have in the perceived audio quality when reproduced through different audio formats. To this end, a set of audiovisual scenes adapted to conventional (stereo, 5.1 Surround) and advanced audio systems (7.1 Surround, 10.1 Surround with height and 3D Binaural sound) is evaluated following the proper international recommendations. This assessment provides a formal study of the impact that advanced spatial audio formats have in the perceived audio quality when different types of common content material are considered. The research questions to be addressed are as follows:

- *How dependent are spatial audio perception attributes on the chosen reproduction system within well-defined audiovisual context?*
- *How are these attributes affected by common broadcasting programme material?*
- *Are there significant differences among different sound systems for every type of broadcasting content?*
- *Which genres are more likely to benefit from these audio formats and to what extent?*

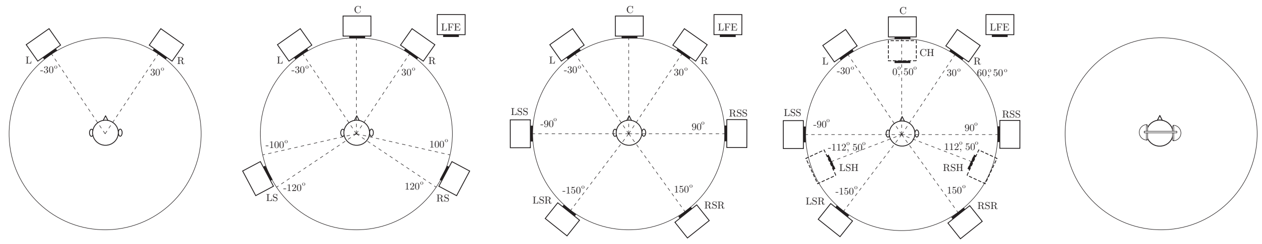
The paper is structured as follows. Section 2 provides a brief introduction to spatial sound and the multichannel audio formats considered in this work. Section 3 describes the background for audiovisual quality evaluation, with emphasis on the international recommendations for the assessment of audio quality within an audiovisual context. Section 4 provides a detailed description of the experimental design issues involving the assessment, including the sound attributes and audiovisual contents evaluated throughout the test sessions. Results are discussed in Section 5 and, finally, conclusions are summarized in Section 6.

2 Multichannel Audio Systems

The objective of spatial sound systems is to accurately recreate the acoustic sensations that a listener would perceive inside a particular environment with certain acoustic properties. This concept, easy to understand, implies a series of physical and technological difficulties that are a current research issue in sound engineering. In this section, we briefly describe the multichannel audio formats evaluated in this work.

2.1 Stereo

Today, the stereo format is still the most common format used for the commercial distribution of sound recordings. The practical experience and a variety of formal



(a) Stereo (b) 5.1 Surround (c) 7.1 Surround (d) 10.1 Surround (e) 3D Binaural

Fig. 1 Multichannel audio formats considered in the evaluation.

research works [28] state that the optimum loudspeaker configuration for stereo is an equilateral triangle with the listener located just to the rear of the point of the triangle as seen in Figure 1(a). Outside this “sweet spot”, phantom images [57](the apparent locations of sound sources in between the loudspeakers) become less stable, and the system is more susceptible to head rotation.

2.2 Surround 5.1

The development of surround sound technology began as early as before the World War II and, from the very beginning, it has been driven by the movie industry. The most known surround system is 5.1, which enables the provision of stereo effects or room ambience to accompany a primarily front-orientated sound stage [51]. Essentially, the three front channels (L, R, C) are intended to be used for a conventional three-channel stereo sound image, while the rear/side channels (LS and RS) are only intended to generate supporting ambience, effects or “room impression”. Figure 1(b) shows the 3-2 format reproduction according to the ITU-R BS.775-1 standard [4]. The “.1” of 5.1 refers to a dedicated *low frequency effects* (LFE) channel or sub-bass channel and it is called “.1” because of its limited bandwidth.

2.3 Surround 7.1

The evolution of 5.1 Surround is the 7.1 Surround format. It is a straightforward extension of 5.1 that adds two additional surround channels (LSS and RSS) at the sides of the listener. Nowadays, many audiovisual productions are distributed in 7.1, since Blu-Ray storage formats provide up to 8 channels of *DTS-HD* lossless audio, *Dolby TrueHD* or 96 kHz/24 bit LPCM audio. The geometry for 7.1 Surround is not yet clearly defined. On the one hand, the ITU-R BS.775 recommends that, in case of using a greater number of surround loudspeakers, these must be symmetrically and uniformly distributed in the arc going from 60° to 150°. On the other hand, Dolby [1] and DTS recommend a configuration where the surround loudspeakers are located at both sides of the listener forming ±90° and ±150° angles with respect to the frontal direction (Figure 1(c)).

2.4 Surround 10.1 with Height

The above 5.1 and 7.1 Surround systems only deliver a horizontal soundfield. In [52], it was reported that listeners preferred sound systems with height. The new generation of surround formats take surround sound to a next level by adding height channels positioned above the basic conventional loudspeaker setup. These elevated loudspeakers are intended to enrich the sound experience with higher depth and dimensionality. Formats such as *10.2 Surround* [34], *22.2 Surround* from NHK [30], *9.1/10.1 Auro3D* [58] and *Dolby Pro Logic IIz* [2] are some of the proposed advanced surround systems with height. The number of elevated loudspeakers varies among these formats, for example, home formats such as Dolby Pro Logic IIz and 9.1 Auro3D use 2 (front) and 4 loudspeakers (front and rear) above the head, respectively. The configuration adopted in this work is shown in Figure 1(d), which has been selected to be a “mean” of the above systems by considering 3 height channels. Two of the elevated loudspeakers were positioned behind the user to give more weight to surround effects, while only one elevated loudspeaker was placed in front of the listener to provide a more stable frontal image. In this work, we refer to this system as 10.1 Surround (with height). Moreover, the use of three elevated loudspeakers has been shown to be sufficient for high-quality surround sound with height [43].

2.5 3D Binaural Sound

In an anechoic environment, as sound propagates from the source to the listener, the different structures of the listeners own body will introduce changes to the sound before it reaches the ear drums [19]. The effects of the listener's body are captured by the *Head-Related Transfer Function* (HRTF), which is the transfer function between the sound pressure that is present at the center of the listener's head when the listener is absent and the sound pressure developed at the listener's ear. While the HRTFs of most humans share many similarities, more detailed examination reveals subtle differences determined mainly by differences in body shape and size among subjects. These subject-dependent differences have been shown to play a major role for precise localization. Binaural sound reproduction is based on the appropriate filtering of anechoic source signals with the HRTFs corresponding to a given spatial direction. It is believed that only using ones own HRTF can result in realistic and accurate binaural audio, as evidenced by various experiments [46]. As opposed to the rest of systems, the sound must be reproduced through headphones to avoid crosstalk effects (Figure 1(e)).

3 Subjective Audio Quality Assessment with Accompanying Video

3.1 Audiovisual Quality

Subjective quality assessment methods are widely used to measure the quality of various audio-visual systems, identifying those critical factors that can cause degradation. This assessment is useful for developing new products and to better understand those aspects that play a key role in the user experience [42]. The principles underlying human perception must be always considered within this context [29]. On the one hand, low-level sensorial processing provides the basic mechanisms for extracting information features from the incoming physical stimuli. On the other hand, high-level cognitive processing sets the basis for the interpretation of quality by merging knowledge, expectations and attitudes into perception. Both processing levels are known to interact between each other, having joint-effects on the final perceived quality [39]. Similarly, the links established between the different sensorial channels making up multimodal perception should not be ignored when studying audio-visual quality. Traditionally, audio and video quality have been studied separately without paying much interest to their interrelationship. As discussed below, this approach neglects the fact that both audio and video are presented together to the user in many final applications. Research on multi-modal perception is of major interest to understand the mutual influence between visual and auditory stimuli as well as to identify those factors that affect the perceived audio-visual quality [33]. First, it must be noted that combining two modalities (e.g. audio and video) is more than the simple sum of two different perceptual channels [31]. Experiments have demonstrated that there is a significant mutual influence between the visual and auditory stimuli. The judgment of the quality in one modality is influenced by the presence of other modality when a combined audio-visual stimulus is presented to a subject [22]. As a result, different types of quality assessment tasks can be defined as a function of the type of stimuli and the focus of the assessment [62]. For example, evaluating audio quality in the presence of audio-only stimuli (audio quality assessment) may be quite different from assessing audio quality in the presence of audio-visual stimuli (audio quality with video assessment). Obviously, audio-visual quality assessment requires the use of both audio and video stimuli.

Although most studies have shown that video quality dominates over audio in the perceived audio-visual quality [38, 22], their relationship is also influenced by other factors [41]. Issues such as temporal synchronization between audio and video [48], usage context [23] or the semantic importance of audio-visual contents are known to have an impact on the perceived quality [41]. For example, audio quality has been shown to be more important than video quality in teleconference sequences [31], as audio conveys most of the information in teleconferencing environments. Despite the relevance of the semantic importance of the contents, there are not many works in the literature covering this topic, probably because semantics is a very subjective concept difficult to integrate into assessment methodologies [62].

Regarding the perceived audio quality with multichannel audio configurations, several research works demonstrate the impact of different loudspeaker setups in the sound quality experienced by a viewer [50]. While some of them cover the evaluation of audio quality in the presence of visual material, most of the investigations are very oriented to theater screens, with little scientific literature devoted to the evaluation of medium-sized TV screens. Choosing an optimal viewing/listening condition is not an easy task [53] since the existing international recommendations can be sometimes contradictory [54]. In this context, meeting auditory and visual requirements simultaneously is not always possible. Section 3.2 discusses the above issue.

3.2 Optimum Viewing & Listening Conditions

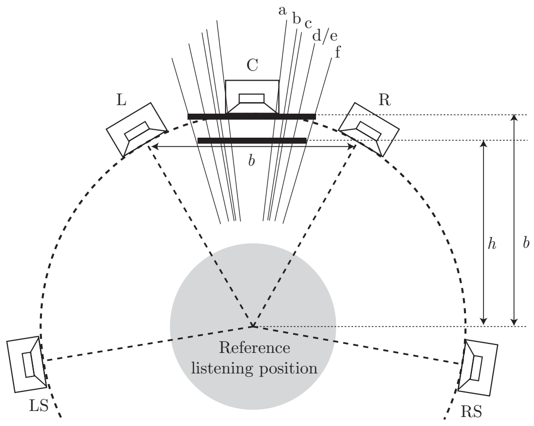
There are different recommendations of the ITU devoted to audio and video subjective quality evaluation [12],[3],[11],[8]. The conditions and attributes to evaluate in these tests are substantially different depending on the relative priority given to the two modalities, with notable differences between the reference room conditions used in both types of tests (video or audio evaluation). For early TV systems (having 4:3 aspect ratio), the recommended viewing distance was six times the height of the screen (denoted as $6H$, being H the height). This recommended viewing distance has been gradually reduced with the emergence of new screen formats, although there is not still a general agreement regarding this point. There are many existing recommendations that refer to the assessment of subjective quality in conventional television systems [7], standard definition digital television [7] or high definition [8]. From the results in [27], it could be extrapolated that the optimum viewing distance is about $2.6H$ for the 1080i video format.

The relationship between viewing distance and loudspeaker distance was one of the key features intended to be covered by ITU-R BS.775-1 Recommendation [4]. The EBU 3276-E [13] also specifies the screen sizes and viewing distances that must be considered when audio and video are jointly presented to a user. The document emphasizes the conflicts that can occur to meet separately the viewing and listening requirements. For example, the presence of a screen causes difficulties in positioning the frontal center loudspeaker in a surround format. Regarding the size of the screen, there are again different possibilities depending on the distance between loudspeakers, the aspect ratio and the viewing distance. Table 1, extracted from this recommendation, shows some of the possibilities based on these parameters. Note that the parameters b and h are defined as in Figure 2.

3.3 Recommendations for the Subjective Assessment of Audio Quality

The ITU-R BS.1283 Recommendation [5] provides a description of the documents governing the subjective assessment of sound quality in different application scenarios. In general, the methods used for the subjective assessment of sound quality itself and the performance of audio systems depend to some extent on the intended purpose of the assessment:

- *Recommendation ITU-R BS.1284* [12] is focused on the general assessment of the quality of sound. It refers to Recommendation ITU-R BS.1116 which contains common requirements.
- *Recommendation ITU-R BS.1116* [3] is the most critical. It is aimed at evaluating systems that introduce very small impairments. The ITU-R emphasizes that this recommendation is not very appropriate when evaluating systems having easily identifiable differences, since it may lead to less reliable results than those obtained by employing a simpler test method. In any case, the ITU-R BS.1116 forms the baseline of other Recommendations, which tend to relax the conditions required by the BS.1116.
- *Recommendation ITU-R BS.1285* [6] is focused on the preparation and preliminary screening of audio systems that are intended to be more strictly evaluated according to the ITU-R BS.1116. This preliminary analysis obviates the need to test those systems that introduce very noticeable differences or impairments.
- *Recommendation ITU-R BS.1286* [9] covers those aspects of the subjective assessment that are particularly relevant in the case where the sound is accompanied with related pictures. Some aspects of the perceived sound quality are influenced by the accompanying visual material.



Code	TV Standard	Aspect Ratio	Viewing Distance	Viewing Angle
a	SDTV	4:3	$6H$	13
b	SDTV wide	16:9	$6H$	17
c	SDTV	4:3	$4H$	19
d	SDTV wide	16:9	$4H$	25
e	HDTV	16:9	$4H$	25
f	HDTV	16:9	$3H$	33

Fig. 2 Viewing angles for different television systems. Reproduced from [13].

4 Experimental Test Design

After reviewing the above international Recommendations and taking into account our specific research context, the ITU-R BS.1286 [9] is selected as the reference document for evaluating the multichannel audio formats described in Section 2.

4.1 Generalities and Objective

Audio and video are inseparably combined in television, movie theaters and other multimedia applications.

Table 1 Relationship between some viewing and listening arrangement parameters [13]

	Viewing distance b					Viewing distance h				
Aspect Ratio	16:9 (Widescreen, HDTV)			4:3		16:9 (Widescreen, HDTV)			4:3	
Viewing Distance (multiple of H)	3	4	6	4	6	3	4	6	4	6
Screen Width (fraction of b)	0.59	0.44	0.30	0.33	0.22	0.51	0.38	0.26	0.29	0.19
Screen Width (meters)	1.19	0.89	0.59	0.67	0.44	1.03	0.77	0.51	0.58	0.38
Screen Diagonal (inches)	54	40	27	33	22	46	35	23	28	19

* b and h refer to Fig.2

** Recommended for HDTV in ITU-R Recommendation BT.710-3

*** Based on a listening circle of 2 m radius

Under normal circumstances, observers should perceive sound and images with a sense of unity. Therefore, the presentation of visual material must be indispensable for the evaluation of some aspects of sound quality in most audiovisual applications. As already explained, visual stimuli normally affect the perception of sound. For example, the apparent direction of a sound source is usually related to its corresponding image, an effect commonly known as the "ventriloquism effect" [32]. In addition, visual stimuli sometimes make small sound impairments and nuances harder to perceive. In this sense, the evaluation of the subjective sound quality accompanied with image must consider several aspects that are of particular interest [9]:

1. The correlation between image and sound.
2. The influence of the presence of visual stimuli on the perceived audio quality.
3. The consistency of the spatial impression evoked by visual and auditory cues.
4. The assessment of the viewing and listening settings.

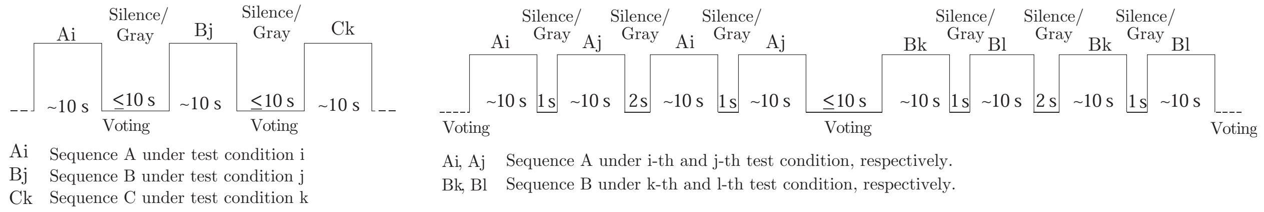
The general objective of the test is to valuate the advantages and disadvantages offered by advanced multichannel audio systems in common broadcasting audiovisual programmes. In this context, besides studying the attributes related to overall sound quality, priority is given to the study of those specific aspects associated with the corresponding visual material for representative broadcasting contents.

4.2 Design Considerations

For the experimental design, issues highlighted in the ITU-R BS.1116 Recommendation [3] are considered. A careful experimental design and approach are necessary to ensure that uncontrolled factors do not contaminate the listening tests, so that there are no ambiguities in the results. For example, if the sequence of sound items to assess is identical for all the subjects performing the test, one might think that the answers given by the subjects could be influenced by the chosen sequence rather than by the small differences between items. In addition, non-uniformities in the test conditions must be carefully addressed so that important issues are taken into account in the presentation phase. For example, when the difficulty level of the material to assess changes, the stimuli presentation order should be randomly distributed both within the same session and other subsequent sessions. Furthermore, tests should be designed so that subjects do not fatigue to the point that the accuracy of their answers decrease. Typically, control conditions include the presentation of unimpaired audio materials among the test items. The differences between the ratings given to these control stimuli and those with potential impairments allow to conclude that the given assessments really correspond to an evaluation of such impairments. The use of a reference in our application context is not entirely clear, since it is very difficult to establish an absolute excellent listening condition or system. For example, although binaural reproduction could lead to a more precise location of sound sources, the need for headphones could influence the quality rating.

4.3 Panel Selection

For the selection of listeners, the ITU-R BS.1284 Recommendation [12] is followed. According to this document, expert listeners are preferable to non-experts. From the ITU perspective, expert listeners can be defined as members of a desired sample population with experience in subjective testing. Experts are able to describe an auditory event in detail and are able to separate different events based on specific impairments. They are able to describe their subjective impressions



(a) Item presentation in ACR Test

(b) Item presentation in PC Test

Fig. 3 Item presentation in ACR and PC tests (reproduced from [10]).(a) ACR test. (b) PC test.

in detail and they have a background in technical implementations of the systems under test, having detailed knowledge of the influence of particular implementations on subjective quality.

Despite the fact that non-expert listeners might be more representative of the general population and experts are often too critical, it has been found that many non-expert listeners become experts when exposed to prolonged system impairments. Therefore, the use of expert listeners gives a more reliable and faster indication. Moreover, if the evaluated systems are intended to broadcasting applications, the recommendation always suggests the use of expert listeners.

The minimum required number of expert listeners is 10. After a personal interview, we selected a set of 16 expert listeners (11 male and 5 female with ages going from 23 to 41) familiarized with audio processing and evaluation methods (researchers and master students). The selection was motivated by their age, gender, background on subjective testing methodologies and interest in sound and music. Moreover, pure tone audiometry tests were performed for pre-screening purposes by using standardized audiometers. Only one subject had a a hearing threshold above 15 dB HL.

4.4 Test Methods and Rating Scales

When the expected subjective differences between systems are very small, the ITU-R BS.1286 recommends using the double-blind triple stimulus with hidden reference method (described in ITU-R BS.1116). However, when the subjective differences are not so small, the methods described in the ITU-R BS.1284 are preferred. Since the systems to be evaluated present very different spatial features, the assessment methods defined in ITU-R BS.1284 are considered. Additional considerations contained in the ITU-T P.911 [10] are included.

4.4.1 Absolute Category Rating (ACR)

In tests of Absolute Category Rating the test sequences to be evaluated are randomly presented one by one and are independently scored. After each presentation, subjects are asked to assess the quality for each of the attributes in Section 4.5 following the next scale:

- 5 - Excellent
- 4 - Good
- 3 - Fair
- 2 - Poor
- 1 - Bad

The method provides no explicit reference condition. The time sequence used in the presentation of the stimuli is shown in Figure 3(a), as recommended by the ITU-T P.911 [10]. The voting time should be less than or equal to 10 s, depending on the voting mechanism used. The presentation time may be somewhat higher or lower depending on the contents of the test sequence (in our case, all the test sequences had equal duration).

4.4.2 Pair Comparison (PC)

In pair comparison tests, sequences are randomly presented in combinations of two elements, where each sequence is first presented via a system under test and subsequently under other system under test. Systems under test (A, B, C, etc.) are generally combined in all possible forms obtaining AB, BA, CA, etc. All possible pairs of sequences should be shown in all possible orders (eg, AB, BA). After the presentation of each pair, the subject gives an assessment of which element of the pair prefers attending a particular attribute. In addition, after presenting each pair, a repetition will be played, leaving a small time slot of approximately 1.5 seconds. An outline of the temporal structure of this test is shown in Figure 3(b). The scale to be used is:

- 3 - First much better than Second
- 2 - First better than Second
- 1 - First slightly better than Second
- 0 - First equal to Second
- -1 - First slightly worse than Second
- -2 - First worse than Second
- -3 - First much worse than Second

PC tests are very interesting because the assessment methodology requires a simpler cognitive task than ACR, involving only the comparison of two stimuli against each other. This is certainly easier than assigning an absolute rating to the stimuli without any reference. Moreover, comparison tests are known to be robust and reflect closely the perceived sensations on a psychological scale [59]. PC tests were carried out to study how a given attribute might be differently perceived when comparing one specific audio system against another, complementing the results extracted from ACR tests.

In both ACR and PC tests, the results corresponding to each subject must be normalized with respect to the mean and standard deviation as described in the ITU-R BS.1284. To avoid fatigue effects, each subject performed the test in 3 sessions of 20 minutes and a training phase of 10 minutes. In the first session the subjects performed the ACR test, while the PC test was conducted during the second and last sessions.

4.5 Perceptual Attributes

The perceptual attributes evaluated by the subjects are the ones defined by the ITU-R BS.1116 and the ITU-R BS.1286. These attributes are described as follows:

- **Frontal sound image quality (FSIQ)**: This attribute is related to the localization of the frontal sound sources. It includes source image quality and losses of definition.
- **Impression of surround quality (ISQ)**: This attribute is related to spatial impression, ambience, or special directional surround effects.
- **Correlation of source positions derived from visual and audible cues (CSP)**: This attribute evaluates the correct and positive relationship between the perceived location of visual elements and their corresponding sound.
- **Correlation of spatial impressions between sound and picture (CSI)**: This attribute is related to the expected correspondence between the spatial impressions derived from auditory and visual stimuli.
- **Basic Audio Quality (BAQ)**: This single, global attribute is used to judge all the aspects that lead to a general impression of the overall perceived audio quality.

It should be emphasized that the subjects were instructed to assess the sound quality in association with the video presentation, rather than to assess the sound quality alone.

4.6 Training

The training was carried out by means of a pre-experiment, selected to illustrate the auditory attributes to be evaluated during the actual test. The session was prepared to let the subjects familiarize with the process, including both the testing equipment and the methodology. All the subjects were provided with written instructions. The training sessions were conducted in groups of four subjects. Once all the groups had completed the training sessions, there was a general discussion session to let all the subjects interact and explain how they understood the attributes to be evaluated. The first part of the session was 45 minutes long and it consisted of a group discussion on spatial audio reproduction led by the authors. All the subjects discussed what they understood from the attribute descriptions found in the instructions and their opinion on how these could be influenced by the type of visual content (without doing any listening pre-test). Then, in the second part of the training session, each member of the group was individually trained to get familiarized to the listening environment and the test material. This part was 35-40 minutes long. To let the subjects concentrate on the specific sound attributes, the material was first presented without any visual content, letting the subject select freely the audio reproduction system. Then, the material was presented together with its visual content, this time leading the subject to concentrate on correlation aspects between visual and audible cues. After the listening, the subject discussed with the authors their general opinion on the experiment (difficulty to assess or understand the attributes, suitability of the test material, synchronization issues between audio and video, representativeness of content according to their genre, etc.). The day of the test, the subject was again trained before starting during 10 minutes in a similar way.

4.7 Audiovisual Material

The test sequences were selected to stimulate the perceptual attributes to be evaluated while being consistent with the main objective of the assessment, i.e. to be representative of common audiovisual contents in broadcasting. According to the annual report of Mediametrie "*One TV Year in the World*" [17], sports, fiction and entertainment programs are the leading broadcasting genres. As a result, the following scenes (10 second sequences in 1080p HD video) were assumed to be representative of common television genres having a strong audio component:

Table 2 Processing Tools Used for Audio Format Conversion

	Stereo	**5.1 Surround**	**7.1 Surround**	**10.1 Surround**	**3D Binaural**
Movies	ITU 5.1 Downmix	Original	DTS Neural UpMix©	VBAP	Dolby Headphone©
Sports	Original	DTS Neural UpMix©	DTS Neural UpMix©	VBAP	Dolby Headphone©
Animation	ITU 5.1 Downmix	Original	DTS Neural UpMix©	VBAP	Dolby Headphone©
Music	ITU 5.1 Downmix	Nuendo MixConvert©	Original	VBAP	Dolby Headphone©

– **Movies**: A sequence from "*Pan's Labyrinth*" having background music, frontal and surround audio effects in a gloomy atmosphere. Additional audio effects corresponding to elevated visual elements (flying fairies) were included to stimulate the perception of sound systems with height.
– **Sports**: A fragment of a soccer match "*Real Madrid - F. C. Barcelona*" where a goal is scored. The sequence has both audience ambient sound and commentator's speech.
– **Animation**: A sequence from the animation movie "*WALL-E*" having background music and well-located audio effects at different distances and directions.
– **Music Video**: A sequence of the music video "*Now or Never*" from the artist "*Orianthi*".

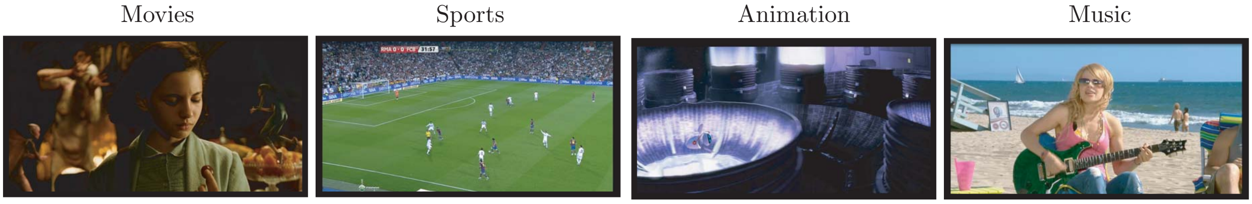
Obviously, having a single 10 s scene for evaluating a content genre is not completely fair, but the nature of the test and the evaluation procedure makes it impractical to include a higher amount of scenes (the required number of combinations and presentation time would become prohibitive). In any case, 93% of the subjects agreed that the selected scenes were enough representative of the above broadcasting genres. A classification of the amount of spatial details, the amount of movement (temporal details) and the amount of scene cuts was performed by using the mean values of 5 expert evaluations. The inter-rater agreement provided by Fleiss' kappa was moderate ($\kappa = 0.664$, $p < 0.01$). Figure 4 shows the resulting description of the test material.

Note also that, while all the scenes were selected for having clear frontal and surround image components, there were not visible speech sources in the selected scenes. First, the music scene already had a high degree of correlation to the video material (especially in the singer's lips movement). Timing aspects and potential asynchrony issues were assumed to be well-represented by this scene. Second, assuming that speech intelligibility may also have an important weight on the evaluated attributes, the authors considered the sports scene to be representative enough for exciting potential intelligibility-related quality factors. Finally, dialogue scenes could focus the attention of the listeners on the actual speaker's message, distracting the subjects from their attribute evaluation task.

An added difficulty in the selection of scenes is the little or null availability of original sequences mixed in all the considered audio systems, since some of them such as 10.1 Surround or 3D Binaural are not standard audio formats. To solve this problem, additional effects, up-mixing processors and specific mixing tools were utilized to adapt the original soundtracks of the scenes to meet the requirements of the systems under test. This adaptation has been done by using state-of-the-art conversion tools belonging to well-known companies in the broadcasting industry such as Dolby and DTS. The use of these tools, besides being interesting from a practical perspective, is coherent with the aim of this work, since the necessity for adapting existing audiovisual material to future broadcasting formats will obviously arise in the next years. Table 2 shows the processors used for the adaptation of the original audio format to be reproduced over the rest of audio systems. These processors are:

– **ITU 5.1 Downmix**: Mixing Matrix specified in the ITU-R BS.775 Recommendation for 5.1 to Stereo Downmix [4].
– **DTS Neural UpMix©** [15]: Processor from DTS oriented to the broadcasting industry for converting Stereo audio to 5.1 Surround or 7.1 Surround.
– **Nuendo MixConvert©** [14]: VST Plug-in from Steinberg for performing spatial audio processing tasks (down-mixing/up-mixing). In this work, it has been used to perform 7.1 to 5.1 downmix.
– **VBAP**: 10.1 mixing has been performed by a specific Matlab tool (there are no commercial processors available) based on Vector Base Amplitude Panning [49].
– **Dolby Headphone©** [18]: Processor from Dolby used to simulate 5.1 recordings through headphones by using HRTF-based techniques.

It must be highlighted that, besides using the above processors, additional audio effects where added to the original soundtracks to stimulate the differences among the different systems. These effects where consistently mixed in all the formats to keep the expected sound source locations unaltered (or as similar as possible). The mixing of these additional effects were performed with Steinberg's Nuendo software. 3D Binaural mixing



Genre	Content	Spat. Detail	Temp. Detail	Scene Cuts	Audio	Length
Movies	Pan's labyrinth	High	Med	Low	Music, FX, Ambience	10 s
Sports	Soccer Madrid-Barcelona	Med	High	Low	Speech, Ambience	10 s
Animation	Wall-E	High	High	Med	Music, FX	10 s
Music	Orianthi - Now or never	Low	Low	High	Music+Vocals	10 s

Fig. 4 Stimuli properties

was performed by means of the VST plug-in *H3D Binaural Spatializer* from *Longcat* [16].

4.8 Equipment and Room Conditions

4.8.1 Audio

The audio playback conditions were controlled to comply with the ITU-R BS.1284 and ITU-R BS.1116 Recommendations. High-quality studio loudspeakers (Dynaudio AIR 6/15 and Dynaudio AIR BASE 1 models) and headphones (Sennheisser HD600) were used in the tests, all of them meeting the requirements of the ITU-R. BS.1116.

4.8.2 Video

As described in Section 3.2, there are several recommendations of the ITU-R that indicate the relationship that should exist between screen size and viewing distance, and the relationship between the loudspeaker setup and the listening distance. The ITU-R BS.1286 recognizes the incompatibility of these recommendations, so it suggests a recommended viewing distance of $3H/4H$ for HDTV and $4H/6H$ for conventional television systems. Recall that H refers to the height of the screen. For the experiments, it was chosen a 42" Full-HD TV ($H = 0.52$ m). Therefore, the appropriate viewing distances should be between $3H$ (1.56 m) and $4H$ (2.08 m). We chose a viewing distance of 1.8 m, placing the speakers over a radius of 2 meters to follow the recommendation. The outline of the configuration of loudspeakers and the listening/viewing area are shown in Figure 5(a).

4.8.3 Listening Conditions

The tests were conducted in the studio of the Institute of Telecommunications and Multimedia Applications, which is acoustically conditioned for a reverberation time of approximately $T60 = 0.25$ s. The geometry of the room is rectangular, with a volume of 96 m^3 and a floor size of 4.5 x 9.1 m. The noise level in the study was kept below 25 dB(A). Loudness equalization was performed to ensure a conformable listening level. Figure 5(b) shows the loudspeakers used for the different audio formats.

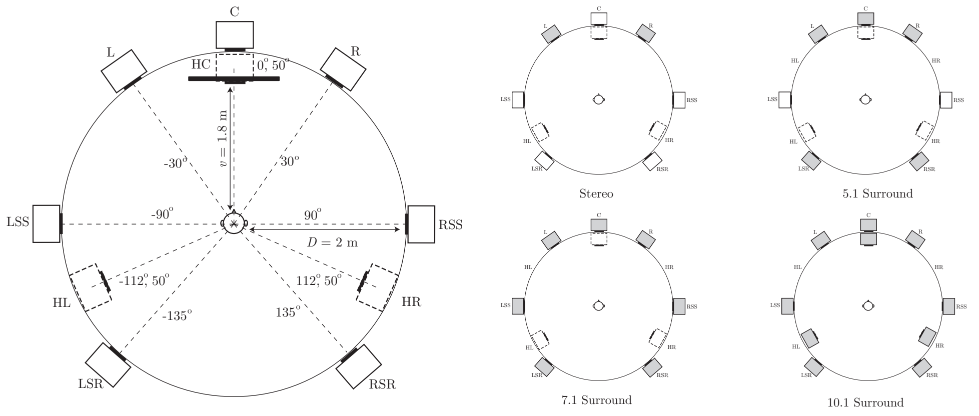
5 Results and Discussion

5.1 ACR Tests

This section presents the results for ACR tests. The results are presented in the form of graphs showing the mean and 95% confidence intervals corresponding to the subjects' responses. Each graph indicates the results for a given content genre, comparing the performance of each audio format according to the attributes explained in Section 4.5.

5.1.1 Movies

Figure 6(a) shows the means and 95% confidence intervals for the movie scene. As expected, the spatial impression in surround systems outperform the stereo format. Although the differences are not excessively high, 10.1 and 7.1 seem to be the best at providing a high spatial impression. However, it is worth to note that the quality of the frontal sound image is slightly better in stereo than in the other systems. Probably, this is due to the fact that subjects are less distracted by surround effects. The worst result in terms of FISQ was for 3D Binaural sound. The typical "inside the head" effect [24] that occurs in HRTF-based systems is probably the explanation for this front image degradation. Regarding the correlation attributes with images, there is a good correlation between sound a visual objects in all the



(a) Loudspeaker Setup

(b) Loudspeaker Activation

Fig. 5 Experimental setup. (a) Loudspeaker setup and listening/viewing area used in the experiments. (b) Loudspeakers used in each audio format.

systems, although surround formats seem to provide a spatial impression more coherent with the visual stimuli. The differences between systems in overall sound quality are not very big, having all of them a score between "good" and "excellent", excluding the case of binaural sound. The reason could be the influence of the discomfort produced by the use of headphones to the listener and the serious lack of power (especially for low-frequency sounds) that usually occurs in headphone reproduction. In any case, the 7.1 Surround system was in average the favorite one, closely followed by 10.1 and 5.1.

5.1.2 Sports

Results for the sports scene are presented in Figure 6(b). In general, the results do not seem to be as favorable as in the case of movies, both in terms of frontal sound image and spatial impression. Note that sound production for sport events is not as thoroughly performed as in the case of movie productions. Sound production in movies requires a lot of time and effort, having usually control over every single item that appears on the screen. In the case of a soccer match scene, the only defined sound source is usually the commentators' voice, being the ambience (audience shouting) the strongest sound component. Moreover, sound production is performed live and the process does not allow any independent treatment of sound sources. The difficulty to perceive a clear position of sound sources is highlighted in the visual/auditory correlation attributes. Although the sense of envelopment is in general lower than in the case of the movie scene, there appears to be a preference for surround systems, in particular 5.1 and 10.1. This preference is also observable in the BAQ attribute.

5.1.3 Animation

Figure 6(c) shows the means and 95% confidence intervals for the animation sequence. A clear preference for 10.1 can be observed, both in terms of frontal image quality and ISQ. Furthermore, there is a considerable improvement in the score for 3D Binaural with respect to other scenes. This sequence has a lot of effects and height source movements, as well as many other distance effects. This might be a good reason for the observed preference of audio formats with height. It is worth to note that making a good use of audio production tools to stimulate the capabilities of audio formats might be decisive in the quality perceived by a viewer. This influence is also marked on the image correlation attributes, since binaural sound and 10.1 got also the best scores. Regarding the perceived BAQ, 10.1 Surround has a better score than the other systems, probably as a result of the factors discussed above.

5.1.4 Music

Results for the music video sequence are shown in Figure 6(d). As with the animation sequence, 10.1 Surround with height was the preferred audio system. Again, this preference seems to be motivated by the enhanced spatial impression, although its score is also slightly better in terms of FSIQ. Although this scene did not include additional audio effects or music instruments (just the original music piece), three new audio tracks were extracted from the original 7.1 mix to create the new 10.1 mix. The new tracks were played through the elevated loudspeakers. The results suggest that, by using

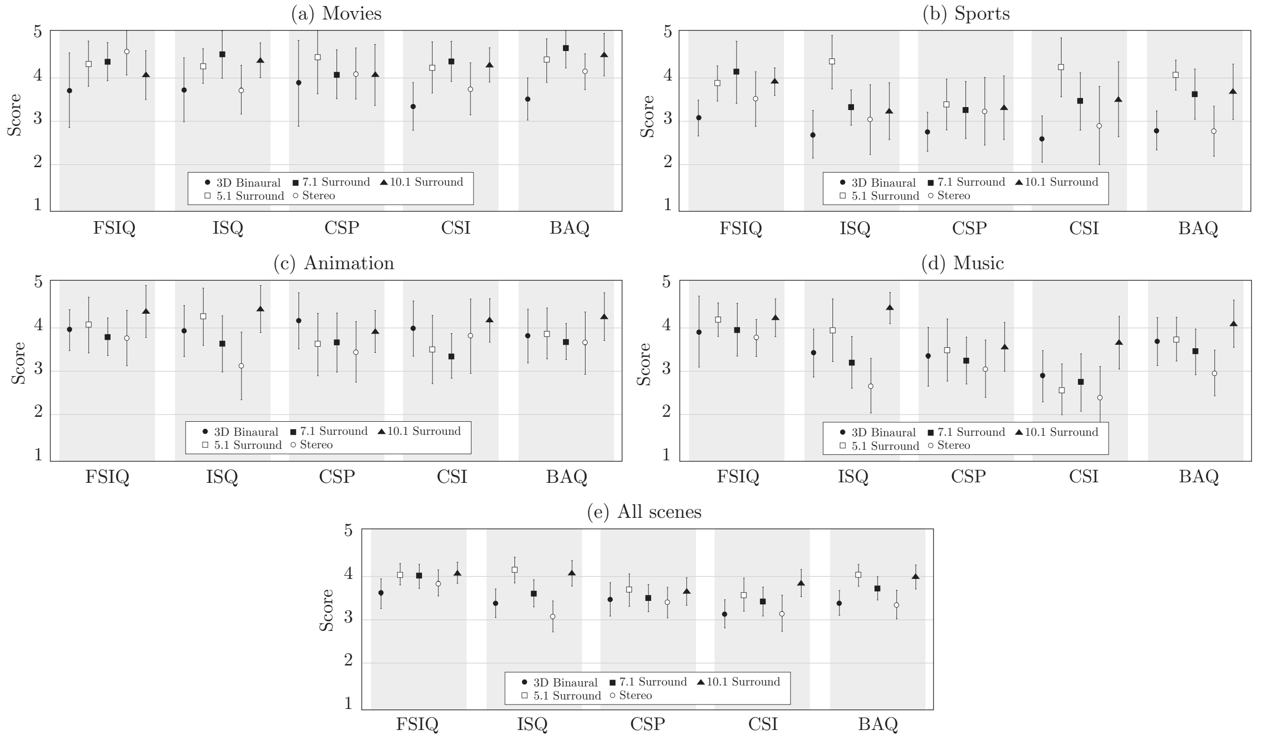
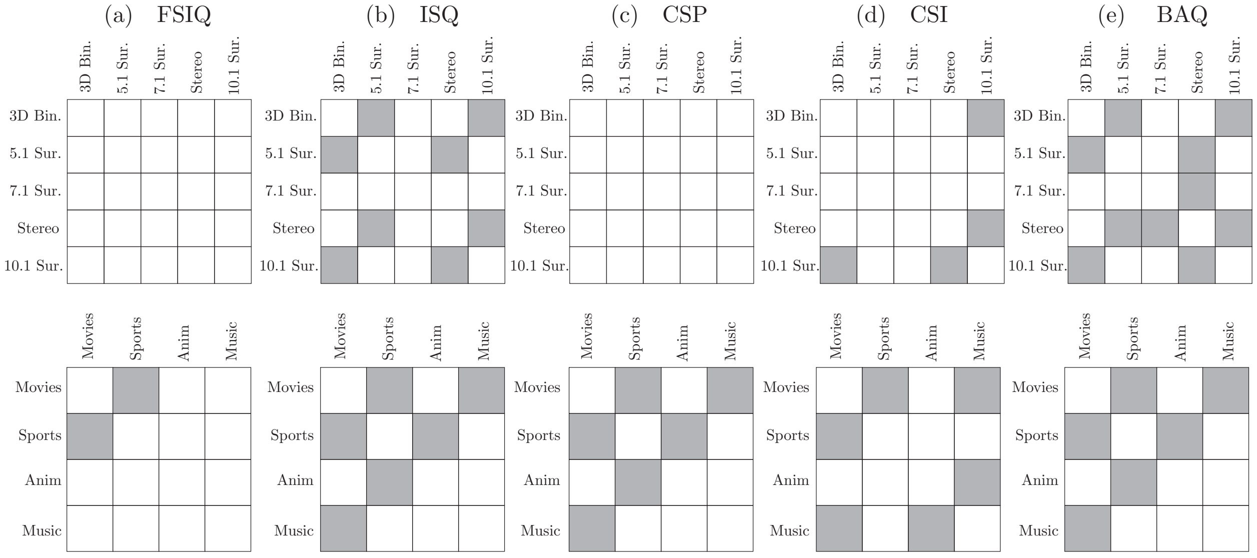


Fig. 6 Results of Absolute Category Rating tests for the different content genres. Bullets denote the mean values for each system under test and bars their corresponding 95% confidence intervals.



(a) FSIQ

	3D Bin.	5.1 Sur.	7.1 Sur.	Stereo	10.1 Sur.
3D Bin.					
5.1 Sur.					
7.1 Sur.					
Stereo					
10.1 Sur.					

	Movies	Sports	Anim	Music
Movies		▨		
Sports	▨			
Anim				
Music				

(b) ISQ

	3D Bin.	5.1 Sur.	7.1 Sur.	Stereo	10.1 Sur.
3D Bin.		▨			▨
5.1 Sur.	▨			▨	
7.1 Sur.					
Stereo		▨			▨
10.1 Sur.	▨			▨	

	Movies	Sports	Anim	Music
Movies		▨		▨
Sports	▨		▨	
Anim		▨		
Music	▨			

(c) CSP

	3D Bin.	5.1 Sur.	7.1 Sur.	Stereo	10.1 Sur.
3D Bin.					
5.1 Sur.					
7.1 Sur.					
Stereo					
10.1 Sur.					

	Movies	Sports	Anim	Music
Movies		▨		▨
Sports	▨		▨	
Anim		▨		
Music	▨			

(d) CSI

	3D Bin.	5.1 Sur.	7.1 Sur.	Stereo	10.1 Sur.
3D Bin.					▨
5.1 Sur.					
7.1 Sur.					
Stereo					▨
10.1 Sur.	▨			▨	

	Movies	Sports	Anim	Music
Movies		▨		▨
Sports	▨			
Anim				▨
Music	▨		▨	

(e) BAQ

	3D Bin.	5.1 Sur.	7.1 Sur.	Stereo	10.1 Sur.
3D Bin.		▨			▨
5.1 Sur.	▨			▨	
7.1 Sur.				▨	
Stereo		▨	▨		▨
10.1 Sur.	▨			▨	

	Movies	Sports	Anim	Music
Movies		▨		▨
Sports	▨		▨	
Anim		▨		
Music	▨			

Fig. 7 Significance matrices for audio systems and genres in ACR tests according to the evaluated attributes. Shaded cells denote significant differences as evaluated by the Tukey-Kramer method.

these elevated speakers, a significant improvement in sound envelopment is produced. Correlation attributes did not get a high score. In fact, music videos tend to be edited so that there is no spatial correlation between sound sources and their corresponding visual objects. This means that many of the images do not show the different performers playing in a well-defined location, but just the main artist performing in different situations or environments. This would explain the low score of all the systems under test regarding correlation with picture. Nevertheless, note that the lack of correlation does not seem to affect the BAQ rating in this genre, probably because most subjects were already used to this issue.

Table 3 ANOVA Results for ACR Tests

	FSIQ		ISQ		CSP		CSI		BAQ	
	F	p	F	p	F	p	F	p	F	p
Genre	2.50	0.0593	6.44	0.0003	9.23	<0.0001	12.25	<0.0001	9.93	<0.0001
Audio System	2.16	0.0736	9.65	<0.0001	0.68	0.605	3.48	0.0084	6.67	<0.0001
Interaction	0.97	0.4769	1.91	0.0332	0.55	0.8792	2.03	0.0217	1.64	0.00792

5.1.5 Average Performance

Figure 6(e) shows the average performance over the different genres. It can be observed that the differences among systems are not as pronounced in terms of frontal sound image quality as in the case of spatial impression. In general, 5.1 Surround and 10.1 are the ones that provide a higher spatial envelopment, followed by 7.1 Surround and 3D Binaural sound. This preference also occurs in terms of correlation attributes, where 10.1 stands as the most capable of generating a sound space more in line with the content presented on the screen. A similar trend is observed with basic audio quality, since 5.1 Surround and 10.1 do also achieve the highest scores.

5.1.6 ANOVA Results

In order to study the interaction between the evaluated audio systems and content genres, a two-way ANOVA analysis was performed. Table 3 contains the values of F and p for the factors “content genre” and “audio system”, as well as their possible interaction effects. The p probabilities obtained for FSIQ are considerably high, which means that it can not be concluded that the changes in the means are actually due to the different content genres or audio systems. This is not the case of ISQ, which shows very small values for p in all the studied factors. Therefore, it can be concluded that the perceived surround quality is very dependent both on the audio reproduction system and the type of content, with a high chance of interaction between them. These interaction effects might be given by the fact that some content genres tend to have a more pronounced ambient component that the others, enhancing the differences between the evaluated audio systems. Regarding the audiovisual correlation of source positions in CSP, it seems that the differences observed through the different genres are significant, not being so the observed among reproduction systems. Moreover, there is not any interaction between the two factors regarding this attribute. However, the p values observed for the correlation of the audiovisual spatial impression are considerably lower, showing a similar behavior as ISQ. Finally, the differences in basic audio quality are both significant at the genre and audio system level, with high probability of interaction between both factors. The next subsection provides deeper insight into all the described effects by analyzing the significance of the differences found among the specific audio systems and genres.

5.1.7 Multiple Comparison

A multiple comparison based on the Tukey-Kramer [44] method has been carried out to examine the significant differences between the different audio systems and genres with respect to the evaluated attributes. Figure 7 shows significance matrices for each attribute, where a shaded cell denotes a significant difference between the items in the corresponding row and column. For ISQ, there are no significant differences among the different audio systems, and only a significant difference between the movie and sports scenes. As expected, surround quality does present more significant differences. The differences found between 3D Binaural and surround systems (5.1 and 10.1) are significant. In fact, these two surround systems are as well significantly different from stereo, while 7.1 Surround does not present a significant difference with respect to any of the other systems. Regarding ISQ, the content genres do also present some significant differences, being movies and sports the ones that present more significant differences. In fact, the differences in ISQ are very similar to those in BAQ, both in terms of audio systems and genres. This reflects a high correlation between ISQ and BAQ, suggesting that surround envelopment has a high weight on the perceived quality. Audiovisual correlation attributes present more significant differences in terms of genres than in audio systems. This fact might reflect that audio-video correlation is much more dependent on content production than on the chosen spatial audio reproduction system.

5.2 PC Tests

The results for pairwise comparison tests (PC) are discussed in this section. Again, the results are presented in the form of graphs showing the mean and 95% confidence intervals of the responses of the subjects. Note that the results are presented using combinations of the systems under test (the markers representing the different audio formats are now presented in pairs at the mean value of the responses).

5.2.1 Movies

Figure 8(a) shows the results for the movie scene. As observed with ACR tests, the frontal sound image quality turns out to be slightly worse in 3D Binaural reproduction than in other systems. This might be again a consequence of the "inside the head" effect. In any case, the differences among other systems regarding FSIQ do not seem too significant. Much greater are the differences found in spatial impression where, as expected, all surround systems (specially 10.1) outperformed stereo. Regarding visual correlation attributes, differences among systems were no very significant in localization but slightly more evident in spatiality (especially with respect to stereo). Finally, BAQ for 5.1 and 7.1 Surround systems is better than for stereo and 3D Binaural, being 10.1 the preferred one.

5.2.2 Sports

PC results for the soccer match scene are in Figure 8(b). The trend in preference for the FSIQ and ISQ attributes is very similar to the observed in the movie scene. Surround systems are slightly better scored than 3D Binaural, although biggest differences are found with stereo. It is worth to note that very little difference is found between 5.1 and 7.1 with respect to 10.1, so having two more additional channels on the horizontal plane does not seem to produce a very strong effect in this scene. Correlation attributes are especially differentiated for 3D Binaural. It is interesting to see that CSI is again higher in the case of 10.1, 5.1 and 7.1. The overall sound quality does not seem to be very different for 5.1, 7.1 and 10.1, although these systems have shown to perform better than 3D Binaural or stereo.

5.2.3 Animation

Figure 8(c) shows the results for animation. This scene does not present very big differences between systems in frontal sound image quality. Again, the greatest differences occur in surround quality, with a clear superiority of 10.1, followed by 7.1 and 5.1. There is a substantial improvement of 3D Binaural in this scene (also noted for ACR tests), probably due to the higher amount of audio effects having a clear location. The highest score in picture correlation attributes is for 10.1 Surround, both in CSP and CSI. The same happens for BAQ, where 10.1 is also the favorite, followed closely by 7.1 and 5.1.

5.2.4 Music

The results for music are in Figure 8(d). The 5.1 system received a very high score, especially in surround audio quality. The rest of surround systems are not very far from this score, although 5.1 seems to be slightly preferred also in frontal image quality. The correlation between visual and auditory cues is somewhat better on surround systems than in stereo or 3D Binaural, although the latter seems to be preferred at most attributes. Regarding overall sound quality, surround 10.1 and 5.1 are preferred, although the results for stereo and binaural are slightly better than in other genres, probably because users are more used to listening to music through headphones.

5.2.5 Average Performance

Figure 8(e) shows the averaged responses over all the scenes. It can be observed that there is a preference for surround systems regarding FSIQ and ISQ, being more pronounced in the latter attribute. Stereo seems to provide similar FSIQ than 5.1 and 7.1, but the ISQ is highly enhanced in surround formats (especially in 10.1). The differences between surround formats and stereo are not so big for picture correlation attributes. The averaged BAQ shows a preference for 10.1, although the scores obtained by 5.1 and 7.1 are not very far. In all cases, the preference for surround formats and binaural with respect to stereo is quite clear.

5.2.6 Preference Analysis

To analyze the preference of the different audio systems from the obtained paired-comparisons, the Thurstone-Mosteller least squares method was used [47][60]. To this end, although the tests were performed by considering a scale of different degrees of preference, the method only takes the number of times that a given system has been preferred over another. Figure 9 shows how the different audio systems are placed on a preference scale for the evaluated attributes and content genres. In general, the results show a preference of 10.1 surround with respect to the other systems, while stereo and 3D Binaural are the least preferred ones. The other surround systems (5.1 and 7.1) are very close in preference, usually somewhere inbetween 10.1 and binaural. As in the case of ACR tests, the distance between audio systems tends to be greater in attributes such as ISQ and BAQ, while FSIQ and CSP tend to show fewer differences in preference. Moreover, the preference order changes depending on the content genre for some attributes. For example, the FSIQ provided by stereo differs in preference over other systems depending on the selected

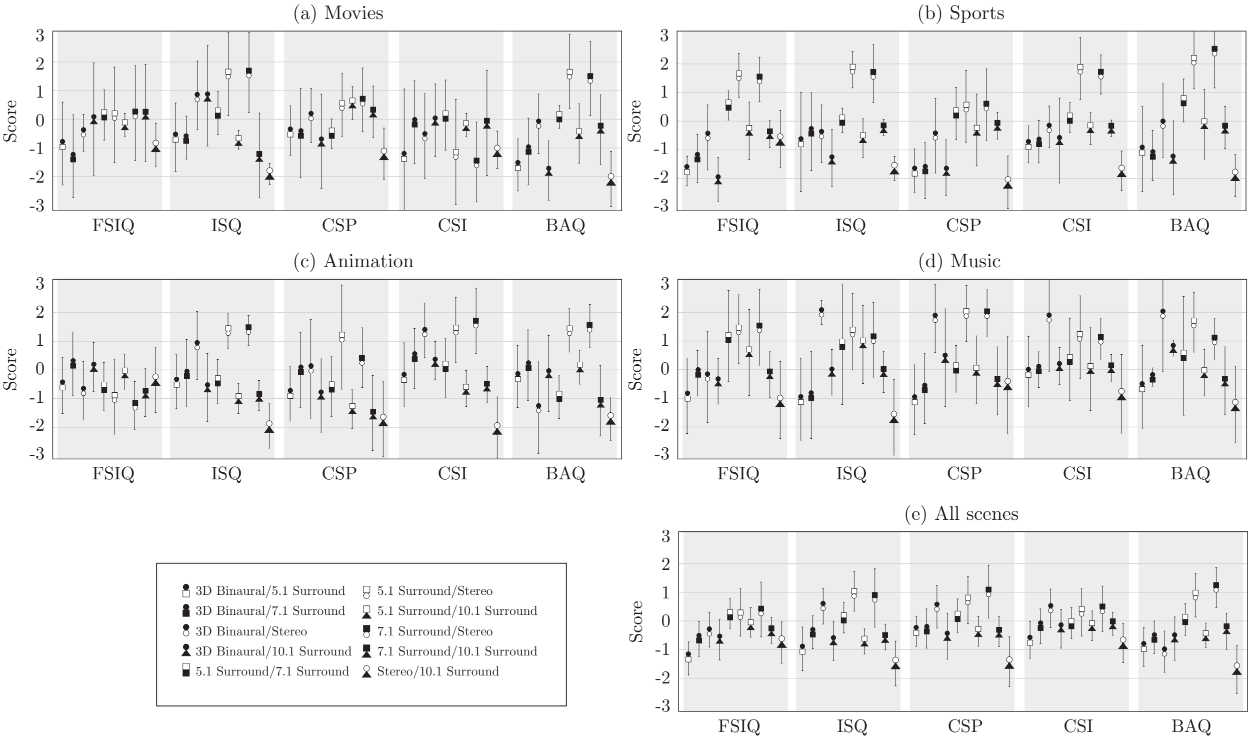


Fig. 8 Results of Pair Comparison tests for the different content genres. Bullet-pairs denote the mean values for each system combination and bars their corresponding 95% confidence intervals.

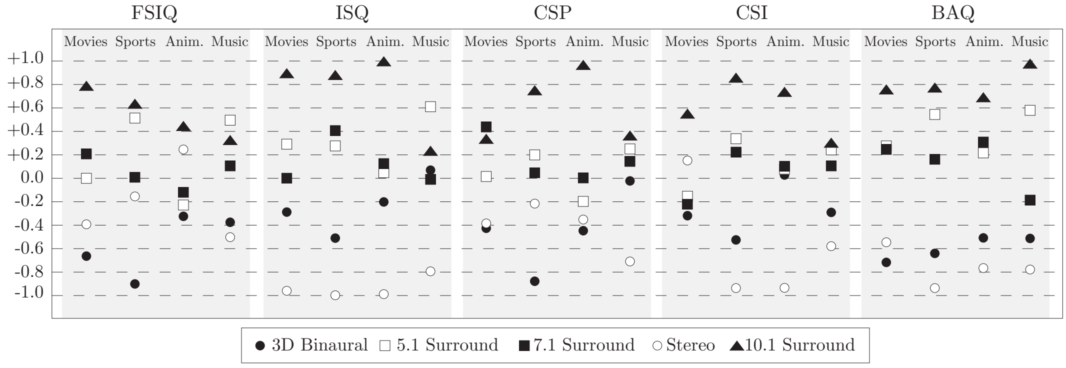


Fig. 9 Audio system preference on an interval scale.

content. Note that the order of preference in important attributes such as BAQ and ISQ has a high correlation with the ratings obtained by means of ACR tests. As a conclusion, both results are quite consistent.

6 Conclusion

This paper presented a formal evaluation of the subjective quality achieved by diverse multichannel audio formats accompanied with video. The objective of this assessment was to analyze which are the benefits of using different multichannel audio formats when the sound accompanies common broadcasting audiovisual contents such as movies, sports, animation or music videos. This knowledge allows to better understand how audio reproduction techniques influence the perception of sound as a function of the specific audiovisual content and which will be the audio needs arising in future immersive broadcasting applications. To this end, a set of representative audiovisual scenes were mixed in the considered audio formats (Stereo, 5.1 Surround, 7.1 Surround, 10.1 Surround with height and 3D Binaural) by using specific spatial audio processing tools. Absolute category rating and pair comparison tests have been conducted following the international recommendations.

On the one hand, results have shown that, in general, the type of audiovisual content has a big influence on the perception of the studied sound attributes. One of the reasons that might affect the perceptual differences among genres is the type of audio production methodologies followed in content production. While some genres such as movies have a thorough audio production stage that allows for a better use of surround capabilities, there are others that rely strongly on live audio production (e.g. sports). In this context, it is quite difficult to assess which are the relative weights that audiovisual interaction issues and content production methodologies have on the perceived attributes. According to the results, the most significative differences among genres were found to be on surround audio quality, audiovisual correlation and basic audio quality. Specifically, movies and animation were significantly different from sports and music. Note that movie and animation contents are more likely to excite surround-related attributes and, as also suggested by the results, surround quality seems to have a high weight on the perceived overall quality. The correlation between audio and video objects was also shown to be very dependent on the particular content type, being better perceived in genres with localized objects (movies and animation).

On the other hand, the type of spatial reproduction system seems to have a relevant impact only on surround and basic audio quality. Significant differences were found between surround audio formats and two-channels systems (Stereo and 3D-Binaural). Moreover, the interaction probabilities between genres and audio systems obtained from the ANOVA analysis show that, as expected, the ratings for these two attributes (surround quality and basic audio quality) are probably mutually influenced. This would confirm that the type of content has a big influence on the excited surround-related attributes, which are more likely to be properly enhanced by multichannel audio systems.

Regarding the preference of the different audio systems, 10.1 Surround was consistently preferred over other systems, usually followed closely by 5.1 and 7.1 Surround formats. Interestingly, paired-comparison tests show a bigger difference than ACR in the preference between 10.1 and 5.1/7.1. Moreover, the relative differences between the preferences for 5.1 and 7.1 are generally much smaller than the ones with respect to 10.1 Surround. This suggests that having elevated loudspeakers has a stronger impact on the perceived quality than the use of additional channels in the horizontal plane. The performance of 3D Binaural sound, despite being a very sophisticated reproduction method, was not as good as expected. This was probably motivated by typical problems such as the "inside the head" effect, headphone discomfort and the lack of bass power.

Finally, it must be highlighted that special care must be taken when generalizing the results obtained in this work. The test material used in this study, while selected for being representative of the considered genres, only represents a very small fraction of possible content types and audiovisual production techniques.

Acknowledgements The Spanish Ministry of Economy and Competitiveness supported this work under the projects TEC2012-37945-C02- 01/02 and IMMERSIVETV.

References

1. Dolby 7.1 home theater speaker guide. available on-line at http://www.dolby.com/. Last viewed 05/07/12
2. Dolby ProLogic IIz. available on-line at http://www.dolby.com/. Last viewed 05/07/2012
3. Recommendation ITU-R BS.1116-1: Methods for subjective assessment of small impairments in audio sysems including multichannel sound systems (1994)
4. Recommendation ITU-R BS.775-1: Multichannel stereophonic sound system with and without accompanying picture (1994)
5. Recommendation ITU-R BS.1283: Subjective assessment of sound quality - a guide to existing recommendations (1997)
6. Recommendation ITU-R BS.1285: Pre-selection methods for the subjective assessment of small impairments in audio systems. (1997)
7. Recommendation ITU-R BT.1128-2: Subjective assessment of conventional television systems (1997)
8. Recommendation ITU-R: 710-4: Subjective assessment methods for image quality in high-definition television (1998)
9. Recommendation ITU-R BS.1286: Methdos for the subjective assessment of audio systems with accompanying picture (1998)
10. Recommendation ITU-T P.911: Subjective audiovisual quality assessment methods for multimedia applications (1998)
11. Recommendation ITU-R 500: Methodology for the subjective assessment of the quality of televisiion pictures (2002)
12. Recommendation ITU-R BS.1284-1: General methods for the subjective assessment of sound quality (2003)
13. EBU Tech 3276-E: Supplement 1 - listening conditions for the assessment of sound programme material: Multichannel sound (2004)
14. Nuendo 3: Operation manual (2005). Steinberg Media Technologies, GmbH
15. Neural^{tm} upmix by DTS user guide (2010). DTS Document Number 9302J70400B
16. H3D binaural spatializer manual. Longcat Audio Technologies SARL (2011)
17. One TV year in the world (2011 issue). Tech. rep., Mediametrie (2011)
18. Dolby headphone webpage. http://www.dolby.com/us/en/consumer/technology/home-theater/dolby-headphone.html, last viewed 07/05/2012 (2012)
19. Algazi, V.R., Duda, R.Q.: Headphone-based spatial sound. IEEE Signal Processing Magazine **28**(1), 33–42 (2011)

20. Apostolopoulos, J., Chou, P., Culbertson, B., Kalker, T., Trott, M., Wee, S.: The road to immersive communication. Proceedings of the IEEE **100**(4), 974 –990 (2012). DOI 10.1109/JPROC.2011.2182069
21. Bech, S., Zacharov, N.: Perceptual Audio Evaluation - Theory, Method and Application. John Wiley & Sons, Chichester, UK (2006)
22. Beerends, J.G., de Caluwe, F.E.: The influence of video quality on perceived audio quality and vice versa. Journal of the Audio Engineering Society **47**(5), 355–362 (1999)
23. Belmudez, B., Moeller, S., Lewcio, B., Raake, A., Mehmood, A.: Audio and video channel impact on perceived audio-visual quality in different interactive contexts. In: IEEE International Workshop on Multimedia Signal Processing, 2009. (MMSP '09) (2009)
24. Blauert, J.: Spatial Hearing. The Psychophysics of Human Sound Localization. MIT Press (1996)
25. Breebaart, J., Faller, C.: Spatial Audio Processing: MPEG Surround and Other Applications. Wiley, Chichester, UK (2007)
26. Brotherton, M.D., Huynh-Thu, Q., Hands, D.S., Brunnstrom, K.: Subjective multimedia quality assessment. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences **E89-A**(11), 2920–2932 (2006)
27. Drewery, J.O., Salmon, R.A.: Tests of visual acuity to determine the resolution required of a television transmission system. BBC R&D White Paper WHP 092 (2004)
28. Eargle, J.M. (ed.): AES Anthology: Stereophonic Techniques. Publications of the Audio Engineering Society, New York (1986)
29. Goldstein, E.B.: Sensation and Perception. Wadsworth Publishing (2002)
30. Hamasaki, K., Hiyama, K., Okumura, R.: The 22.2 multichannel sound system and its application. In: Proceedings of the 118th AES Convention. Barcelona, Spain (2005)
31. Hands, D.S.: A basic multimedia quality model. IEEE Transactions on Multimedia **6**(6), 806–816 (2004)
32. Hershey, J., Movellan, J.: Advances in Neural Information Processing Systems, chap. Audio-vision: using audio-visual synchrony to locate sounds. MIT Press, Cambridge, Mass, USA (1999). Pp. 813-819
33. Hollier, M.P., Rimell, A.N., Hands, D.S., Voelcker, R.M.: Multi-modal perception. BT Technology Journal **17**(1), 35–46 (1999)
34. Holman, T.: 5.1 Surround Sound: Up and Running (2nd Edition). Focal Press (2007)
35. Holman, T.: Sound for film and television (3rd Edition). Focal Press (2010)
36. Huang, Y., Chen, J., Benesty, J.: Immersive audio schemes. Signal Processing Magazine, IEEE **28**(1), 20–32 (2011). DOI 10.1109/MSP.2010.938754
37. Huynh-Thu, Q., Barkowsky, M., Le Callet, P.: The importance of visual attention in improving the 3D-TV viewing experience: Overview and new perspectives. IEEE Transactions on Broadcasting **57**(2), 421 –431 (2011). DOI 10.1109/TBC.2011.2128250
38. Jones, C., Atkinson, D.J.: Development of opinion-based audiovisual quality models for desktop video-teleconferencing. In: Proceedings of the 6th International Workshop on Quality of Services (IWQoS 98). Napa Valley, CA (1998)
39. Jumisko-Pyykkö, S.: User-centered quality of experience and its evaluation methods for mobile television. Ph.D. thesis, Tampere University of Technology (2011)
40. Jumisko-Pyykkö, S., Hakkinen, J., Nyman, G.: Experienced quality factors - qualitative evaluation approach to audiovisual quality. In: Proceedings of 19th SPIE Annual Symposium on Electronic Imaging. San Jose, California, USA (2007)
41. Jumisko-Pyykkö, S., Strohmeier, D.: Cognitive styles and visual quality. In: Proceedings of SPIE 8667, Multimedia Content and Mobile Devices (2013)
42. Jumisko-Pyykkö, S., Weitzel, M., Strohmeier, D.: Designing for user experience: what to expect from mobile 3D TV and video? In: Proceedings of the 1st International Conference on Designing Interactive User Experiences for TV and Video (UXTV '08). Mountain View, CA, USA (2008)
43. Kim, S., Lee, Y.W., Pulkki, V.: New 10.2 - channel vertical surround system (10.2 - VSS); comparison study of perceived audio quality in various multichannel sound systems with height loudspeakers. In: Proceedings of the 129th AES Convention. San Francisco, USA (2010)
44. Kramer, C.Y.: Extension of multiple range tests to group means with unequal numbers of replications. Biometrics **12**, 307–310 (1956)
45. Kyriakakis, C., Tsakalides, P., Holman, T.: Surrounded by sound. IEEE Signal Processing Magazine **16**(1), 55 –66 (1999). DOI 10.1109/79.743868
46. Moller, H., Sorensen, M.F., Jensen, C.B., Hammershoi, D.: Binaural technique: do we need individual recordings? Journal of the Audio Engineering Society **44**, 451–468 (1996)
47. Mosteller, F.: Remarks on the method of paired comparisons: The least squares solution assuming equal standard deviations and equal correlations. Psychometrika **16**(1), 3–9 (1951)
48. Nixon, N.F., Spitz, L.: The diction of auditory visual desynchrony. Perception **9**, 719–721 (1980)
49. Pulkki, V.: Virtual sound source positioning using vector base amplitude panning. Journal of the Audio Engineering Society **45**(6), 456–566 (1997)
50. Reiter, U.: Subjective assessment of the optimum number of loudspeaker channels in audio-visual applications using large screens. In: Proceedings of the 28th AES International Conference (2006)
51. Rumsey, F.: Spatial Audio. Focal Press (2001)
52. Silzle, A., George, S., Habets, E.A.P., Bachmann, T.: Investigation on the quality of 3D sound reproduction. In: Proceedings of the International Conference on Spatial Audio (ICSA 2011). Detmold, Germany (2011)
53. Steinke, G.: Surround-sound: Relations of listening and viewing configurations. In: Proceedings of the 116th AES Convention. Berlin, Germany (2004). Paper 6019
54. Steinke, G.: High definition surround sound with accompanying HD picture. In: Proceedings of the International Tonmeister Symposium. Vabaria (2005)
55. Strohmeier, D., Jumisko-Pyykkö, S.: How does my 3D video sound like? - impact of loudspeaker set-ups on audiovisual quality on mid-sized autostereoscopic display. In: Proceedings of the 3DTV Conference (3DTV-CON'08). Istanbul, Turkey (2008)
56. Theile, G.: HDTV sound systems: how many channels? In: Proceedings of the AES 9th International Conference. Detroit, Michigan (1991)
57. Theile, G.: On the naturalness of two-channel stereo sound. Journal of the Audio Engineering Society **39**, 761–767 (1991)
58. Theile, G., Wittek, H.: Principles in surround recordings with height. In: Proceedings of the 130th AES Convention. London, UK (2011)

59. Thurston, L.L.: A law of comparative judgment. Psychological Review **101**(2), 266–270 (1994)
60. Tsukida, K., Gupta, M.R.: How to analyze paired comparison data. Tech. rep., Department of Electrical Engineering, University of Washington (2011)
61. Wang, K., Barkowsky, M., Brunnstrom, K., Sjostrom, M., Cousseau, R., Le Callet, P.: Perceived 3D TV transmission quality assessment: Multi-laboratory results using absolute category rating on quality of experience scale. IEEE Transactions on Broadcasting **PP**(99), 1 (2012). DOI 10.1109/TBC.2012.2191031
62. You, J., Reiter, U., Hannuksela, M.M., Gabbouj, M., Perkins, A.: Perceptual-based quality assessment for audio-visual services: A survey. Signal Processing: Image Communication **25**, 482–501 (2010)
63. Zhang, L., Vazquez, C., Knorr, S.: 3D-TV content creation: Automatic 2D-to-3D video conversion. IEEE Transactions on Broadcasting **57**(2), 372 –383 (2011). DOI 10.1109/TBC.2011.2122930
64. Zielinski, S., Rumsey, F., Bech, S.: Subjective audio quality trade-offs in consumer multichannel audio-visual delivery systems. part i: Effects of high frequency limitation. In: Proceedings of the AES 112th Convention. Munich, Germany (2002)