TECNIACUSTICA MURCIA · 2014

45º CONGRESO ESPAÑOL DE ACÚSTICA
8º CONGRESO IBÉRICO DE ACÚSTICA
EUROPEAN SYMPOSIUM ON SMART CITIES AND ENVIRONMENTAL ACOUSTICS

SUBJECTIVE QUALITY ASSESSMENT OF MULTICHANNEL AUDIO QUALITY IN REPRESENTATIVE AUDIOVISUAL GENRES

PACS: 43.60

Maximo Cobos[1]; Jose J. Lopez[2]; Ana M. Torres[3], Juan M. Navarro[4], Germán Ramos[5]
1 Dpt Informàtica, ETSE, Universitat de València, Av. de la Universitat s/n, 46100, Valencia, Spain maximo.cobos@uv.es.
2 iTEAM, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain jjlopez@dcom.upv.es.
3 Departamento I.E.E.A.C., Universidad Castilla-La Mancha, 16071 Cuencia, Spain ana.torres@uclm.es.
4 Advanced Telecom. Group, San Antonio's Catholic University of Murcia, Murcia 30107, Spain,jmnavarro@ucam.edu.
5 ITACA Institute, Universitat Politècnica de València, Valencia, 46022 Spain, gramosp@eln.upv.es.

ABSTRACT
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 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.

RESUMEN
Los formatos de televisión HDTV y 3DTV están siendo adoptados satisfactoriamente por el mercado de consumo, teniendo un impacto considerable en la presentación de contenidos actuales. Junto a los avances en vídeo, también se ha producido un impulso considerable de los formatos multicanales de audio. Sin embargo, la necesidad de herramientas específicas de producción y configuraciones específicas de altavoces para cada uno de los sistemas existentes parece estar afectando de forma importante la adopción de estos formatos. Además, es bien sabido que la calidad sonora está altamente influenciada por el contexto, donde la interacción multimodal juega un papel importante. En este artículo se presenta una evaluación de la calidad sonora percibida a través de varios formatos de audio (stereo, sistemas surround avanzados y sonido 3D binaural) acompañados por video. Se consideran para ello diversos géneros audiovisuales representativos: deportes, películas, música y animación, evaluando el impacto de los mismos en los atributos perceptuales contemplados dentro de las recomendaciones internacionales.

1 INTRODUCCIÓN

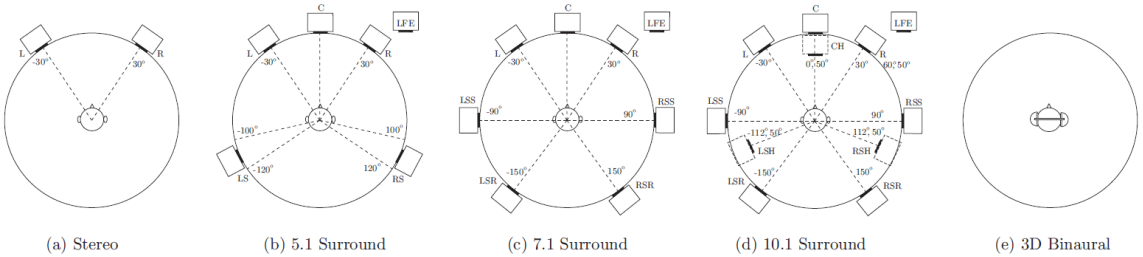
The development of immersive multimedia environments is highly linked to spatial audio reproduction [1],[2]. Stereo sound systems, considered as the simplest approximation to spatial audio, have been utilized throughout the last 80 years as an added value in sound recordings, specially for music material [3]. Together with the entertainment industry, stereo sound evolved to surround sound systems, which provide a better spatial sensation than stereo by using more reproduction channels [4]. 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.

Although the general advantages of using multichannel audio formats in broadcasting seems to be quite clear [5],[6], 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 [7]. However, to the best of the authors' knowledge, there are not previous works focused on the impact that audiovisual content types have on audio perception when conventional and advanced spatial reproduction systems are considered. Although in [8] 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. In this paper, we present a preliminary evaluation of the subjective audio quality provided by several multichannel audio formats accompanied with picture, which is part of a recently completed more extensive study including pair comparison tests [9]. 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 on the perceived audio quality when different types of common content material are considered. The results suggest that, while the complexity of the system in terms of reproduction channels and required processing has generally a big influence in the overall perceived audio quality, the differences perceived among the studied audio formats are very dependent on the reproduced audiovisual content.

2 MULTICHANNEL AUDIO SYSTEMS

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 research works state that the optimum configuration for two-loudspeaker 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) [3].



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

Figure 1. Multichannel audio formtas considered in the evaluation.

2.2.- 5.1 Surround

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. 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 [10].

2.3.- 7.1 Surround

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. Dolby [11] and DTS recommend a configuration where the surround loudspeakers are located at both sides of the listener forming ±90º and ±150º angles with respect to the frontal direction (Figure 1(c)).

2.4.- 10.1 Surround 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. Formats such as 10.2 *Surround* [4], 22.2 *Surround* from NHK [12], 9.1/10.1 *Auro3D* [13] and *Dolby Pro Logic IIz* [14] 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.

2.5.- 3D Binaural

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. The effects of the listener's body are captured by the *Head-Related Transfer Function* (HRTF). Binaural audio is based on creating a realistic spatial experience by filtering the sound sources using a given HRTF model. As opposed to the rest of systems, the sound must be reproduced through headphones to avoid crosstalk effects (Figure 1(e)) [15].

3 EXPERIMENTAL TEST DESIGN

After reviewing the international Recommendations and taking into account our specific research context, the ITU-R BS.1286 [16] is selected as the reference document for evaluating the multichannel audio formats described in Section 2. The evaluation of the subjective sound quality accompanied with image must consider several aspects that are of particular interest such as [16]: the correlation between image and sound, the influence of the presence of visual stimuli on the perceived audio quality, the consistency of the spatial impression evoked by visual and auditory cues and the assessment of the viewing and listening settings. For the experimental design, issues highlighted in the ITU-R BS.1116 Recommendation [17] 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. For the selection of listeners, the ITU-R BS.1284 Recommendation [18] is followed. According to this document, expert listeners are preferable to non-experts. Moreover, if the systems evaluated are intended to broadcasting applications, the recommendation always suggests the use of expert listeners. The mínimum required number of expert listeners is 10. A training sesión must be carried out to let the subjects familiarize with the test procedure, the audiovisual material and

the playback environment. In our case, 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) and verified normal hearing.

3.1.- Test Method and Rating Scale

3.1.1. 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 of the presented sequence following the scale: *5 – Excellent, 4 – Good, 3 – Fair, 2 – Poor, 1 – Bad*. 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 a duration of 10 seconds).

3.2.- 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. The subjects taking part in the tests were explained the meaning of these attributes in a training session preceding the tests. 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.

3.3.- Audiovisual Material

The test sequences (10 seconds long and 1080p) were selected to stimulate the perceptual attributes to be evaluated while being representative of common audiovisual contents in broadcasting:

• *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. 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.

3.4.- Equipment and Room Conditions

The audio playback conditions were controlled to comply with the ITU-R BS.1284 and ITU-R BS.1116 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 3*H*/4 *H* for HDTV and 4 *H* /6 *H* 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 3 *H* (1.56 m) and 4 *H* (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 2(a).

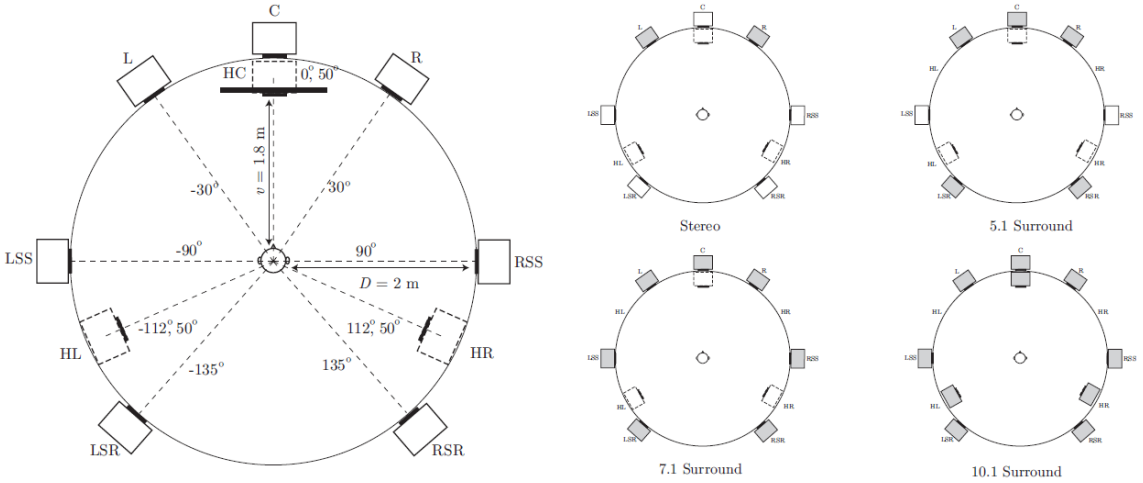


Figure 2. Experimental setup. (a) Loudspeaker setup and listening/ viewing area used in the experiments. (b) Loudspeakers used in each audio format.

4 RESULTS AND DISCUSSION

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 3.2.

4.1.- Movies

Figure 3(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 [19] 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 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. In any case, the 7.1 surround system was in average the favorite one, closely followed by 10.1 and 5.1.

4.2.- Sports

Results for the sports scene are presented in Figure 3(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 require 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.

4.3.- Animation

Figure 3(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.

4.4.- Music

Results for the music video sequence are shown in Figure 3(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 créate the new 10.1 mix. The new tracks were played through the elevated loudspeakers. The results suggest that, by using 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.

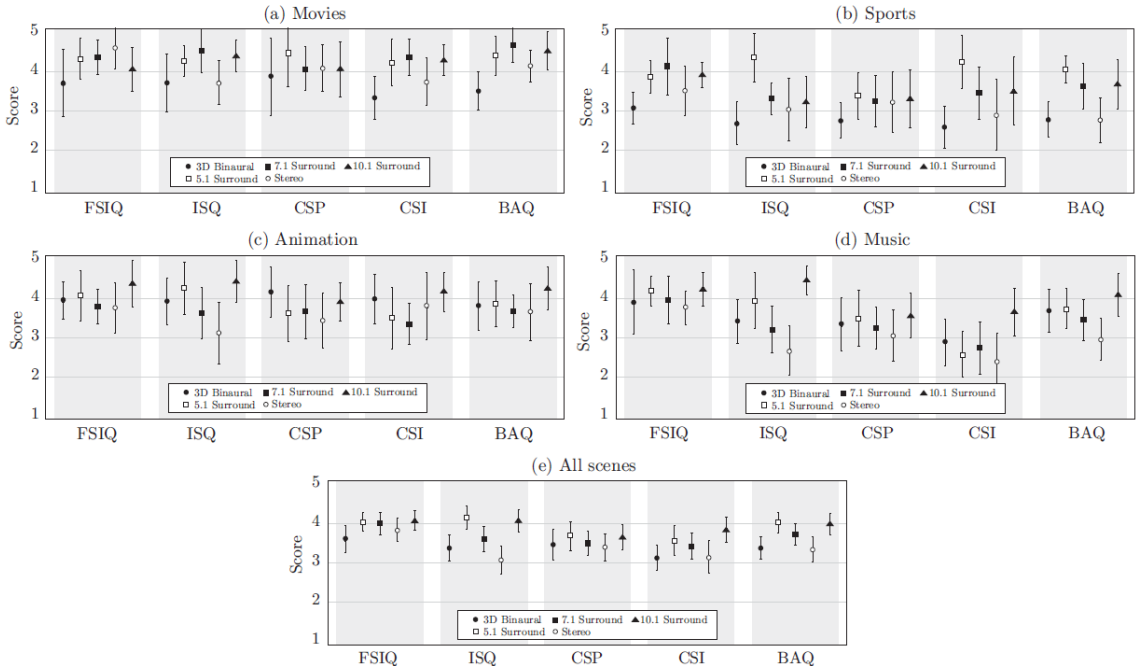


Figure 3. Results of Absolute Category Rating tests for the different content genres. Bullets denote the mean values for each system under test and bars their corresponding 95% confidence intervals.

4.5.- Average Performance

Figure 3(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 CONCLUSION

This paper presented the design and results of ACR tests aimed at evaluating the subjective quality achieved by diverse multichannel audio formats accompanied with video. In this context, audiovisual scenes belonging to common broadcasting genres (movies, sports, animation and music videos) were considered in different multichannel formats: stereo, 5.1 Surround, 7.1 Surround, 10.1 with height and 3D Binaural. Tests have been conducted following the international recommendations and the results have shown that, in general, the type of audiovisual content has a big influence on the perception of the studied sound attributes. While some genres such as movies have a thorough audio production stage that allows for a better use of surround capabilities, the use of more audio channels does not seem to have the expected impact on other types of contents. In this context, movies and animation were shown to be especially favored with 10.1 with height, outperforming 5.1 and 7.1 Surround systems. On the other hand, music and sports did not seem to be specially influenced by elevated channels. Moreover, correlation attributes were shown to be highly dependent on the audiovisual genre, being better perceived in those scenes having well-localized objects. Further details regarding the comparison of multichannel audio format pairs can be found in [9].

6. AGRADECIMIENTOS

Este trabajo ha sido financiado por el Ministerio de Economía y el programa FEDER mediante el proyecto con referencia TEC2012-37945-C02-02.

References

1. C. Kyriakakis, P. Tsakalides, and T. Holman, “Surrounded by sound,” IEEE Signal Processing Magazine, vol. 16, no. 1, pp. 55 –66, jan 1999.
2. F. Rumsey, Spatial Audio, Focal Press, 2001.
3. J. M Eargle, Ed., AES Anthology: Stereophonic Techniques, Publications of the Audio Engineering Society, New York, 1986.
4. T. Holman, 5.1 Surround Sound: Up and Running (2nd Edition), Focal Press, 2007.
5. G. Theile, “HDTV sound systems: how many channels?,” in Proceedings of the AES 9th International Conference, Detroit, Michigan, May 1991.
6. T. Holman, Sound for film and television (3rd Edition), Focal Press, 2010.
7. D. Strohmeier and S. Jumisko-Pyykk¨o, “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 (3DTVCON’ 08), Istanbul, Turkey, May 2008.
8. S. Zielinski, F. Rumsey, and S. Bech, “Subjective audio quality trade-offs in consumer multichannel audiovisual delivery systems. part i: Effects of high frequency limitation,” in Proceedings of the AES 112th Convention, Munich, Germany, April 2002.
9. M. Cobos, J.J. Lopez, J.M. Navarro and G. Ramos, “Subjective quality assessment of multichannel audio accompanied with video in representative broadcasting genres,” Multimedia Systems, vol.19, July 2013.
10. “Recommendation ITU-R BS.775-1: Multichannel stereophonic sound system with and without accompanying picture,” July 1994.
11. “Dolby 7.1 home theater speaker guide,” available on-line at http://www.dolby.com/uploadedFiles/Assets/US/Doc/Consumer/ Dolby-Home-Theatre-Speaker-Guide-7. 1-6-8.pdf, last viewed 05/07/12.
12. K. Hamasaki, K. Hiyama, and R. Okumura, “The 22.2 multichannel sound system and its application,” in Proceedings of the 118th AES Convention, Barcelona, Spain, May 2005.
13. G. Theile and H.Wittek, “Principles in surround recordings with height,” in Proceedings of the 130th AES Convention, London, UK, May 2011.
14. “Dolby ProLogic IIz,” http://www.dolby.com/consumer/technology/prologic-IIz.html, last viewed 05/07/2012.
15. V. R. Algazi and R. Q. Duda, “Headphone-based spatial sound,” IEEE Signal Processing Magazine, vol. 28, no. 1, pp. 33–42, 2011.

16. “Recommendation ITU-R BS.1286: Methdos for the subjective assessment of audio systems with accompanying picture,” 1998.
17. “Recommendation ITU-R BS.1116-1: Methods for subjective assessment of small impairments in audio systems including multichannel sound systems,” 1994.
18. “Recommendation ITU-R BS.1284-1: General methods for the subjective assessment of sound quality,” 2003.
19. Jens Blauert, Spatial Hearing. The Psychophysics of Human Sound Localization, MIT Press, 1996.