IMPROVING SOUND INCIDENCE REPRODUCTION FROM B-FORMAT IMPULSE RESPONSES FOR AURALIZATION OF CONCERT HALLS

PACS: 43.55.Mc

Espitia Hurtado, Juan Pablo[1]; Polack, Jean-Dominique[2]; Warusfel, Olivier[3]
UPMC Univ Paris 06, UMR 7190, Institut Jean Le Rond d'Alembert, F-75005 Paris, France[1,2]
CNRS, UMR 7190, Institut Jean Le Rond d'Alembert, F-75005 Paris, France[1,2]
UMR 9912, Sciences et Techniques de la Musique et du Son, IRCAM-CNRS-UPMC, F-75004 Paris, France[1,3]
E-mail: espitia@lam.jussieu.fr

ABSTRACT

In order to analyze the acoustical properties of fifteen concerts halls and theatres in Paris, listening tests were undertaken using Ambisonics reproduction over twelve loudspeakers. Convolution was performed exploiting the measured first order Ambisonics room impulse responses (B-format RIRs). Results showed, among other weaknesses, a lack of spaciousness that could be linked to a non-optimal sound incidence reproduction. Decoding improvement was achieved by estimating the instantaneous intensity vector and diffuseness of the direct sound and early reflections derived from the RIRs, and by routing the non-diffuse components in the direction of the corresponding intensity vector using VBAP rendering.

RESUMEN

Con el fin de analizar las propiedades acústicas de 15 salas de conciertos y teatros en Paris, se efectuaron pruebas de escucha usando un sistema de reproducción Ambisonics de 12 altavoces. La convolución se llevó a cabo usando las respuestas a impulso Ambisonics de primer orden medidas en las salas (B-Format RIRs). Los resultados mostraron, entre otros puntos débiles, una escasez en la impresión espacial que podría estar relacionada con una reproducción no óptima de la dirección de incidencia del sonido. Una mejoría en la decodificación fue realizada estimando el vector de intensidad instantánea y la difusión del sonido directo y las primeras reflexiones a partir de las RIRs, y orientando los componentes no difusos en la dirección correspondiente al vector de intensidad mediante la técnica VBAP.

1. INTRODUCTION

Auralization is the process of rendering audible the sound field of a source in a space [1]. Thus, auralization has been used for subjective evaluation of concert halls. Contrary to *in-situ* listening tests, auralization allows comparing between different spaces with exactly the same musical source in the same listening conditions. Furthermore, comparisons can be done quickly in time. This is required to listen carefully to the differences between the acoustics of the concert halls

[2]. However, the relevance of the results depends on the degree of fidelity between the real auditory environment and its virtual rendition. In general terms, concert halls auralization is produced by convolving anechoic musical signals with measured spatial room impulse responses (also referred to as directional room impulse response, DRIR). Therefore, the auralization is strongly affected by the choice of the measuring device, the rendering setup and of the encoding and decoding process of these DRIR. A straightforward approach to convey 3D information may consist in recording a binaural RIR using a dummy head and to reproduce the auralization signals on headphones. The main advantage of this technique is that it requires only limited equipment both during the measurement and the listening phases. However, for authentic auralization this method will suffer from perceptual artefacts, such as in-head localisation, linked to the use of a generic dummy head recording which cannot respect the individual spatial cues contained in the listener’s HRTF [3]. More generally, it is important to keep the recorded DRIR format as generic as possible in order to maintain its compatibility with various rendering loudspeaker setups or possibly to allow for its individualized binaural decoding. To this respect, the first-order Ambisonics B-Format or its High Order Ambisonics (HOA) extensions are good candidates [4]. B-Format rendering is spatially homogeneous and is very convenient because of the existence of commercial microphones for recording and also the simplicity in playback/rendering process. However, the image sound is blurred due to poor localization accuracy [5]. In alternative, HOA increases angular discrimination and enlarges the available listening area (the higher the order, the better the spatial resolution [6]). However, HOA requires high spatial resolution microphones (e.g. spherical microphone arrays’) for measuring DRIR, as well as large number of loudspeakers for the decoding.

Other methods have been proposed to exploit B-Format DRIRs using parametric decoding. This is the case for *Spatial Impulse Response Rendering* technique (SSIR) [7], employing sound intensity theory and *High Angular Resolution Planewave Expansion* (HARPEX) [8], based on plane wave decomposition. In both cases, B-Format signals are analysed in time and frequency in order to improve the sound spatial image. Listening tests for both methods were compared with first-order Ambisonics systems showing better results [8, 9]. The SSIR technique has been widely used in concert hall evaluation.

The room acoustics group at Université Pierre et Marie Curie has a data base of B-Format RIRs measured in 2009 in unoccupied concert halls and theatres in Paris selected for their historical, architectural, or acoustical interest. The measurement source was a dodecahedral sound source Outline GRS and a subwoofer Tannoy Power VS10 giving an omnidirectional radiation pattern up to the 8 kHz octave band as imposed in the ISO 3382-1 standard. A 15s exponential sweep-sine from 20 Hz up to 20 kHz was used as excitation signal. The response was measured with a SoundField ST250 microphone. An average of ten microphone positions were used for the three different source positions on stage (center, left and right). Furthermore, between 2010 and 2011, listening tests were also conducted from those measurements using a *basic* first-order ambisonics decoder in a listening room consisting of 12 loudspeakers positioned in dodecahedral form [10]. Results showed, among others weaknesses, a lack of spaciousness that could be linked to a non-optimal sound incidence reproduction.

This paper studies the improvement of the spatial rendering achieved by exploiting the sound intensity theory for decoding the B-format RIRs. The merit of the method is estimated through the comparison of the intensity vector associated to the direct sound and of some conventional acoustical descriptors between the real and reproduction contexts. The decoding method is closed to SSIR, although it is conducted in the time domain and is restricted to the direct sound and early reflections only. Late reflections are rendered by a *basic* B-Format decoder. It is known that early lateral reflections contribute to the spatial impression of halls, as was proven by Barron and Marshall [11]. In addition, Griesinger [12] suggests that if direct sound is clearly distinct, as is the case with accurate localization, it is possible for the brain to separate this perception from the perception of reflections and reverberation and in consequence to perceive a better enveloping sound. For this reason, decoding improvement was achieved by estimating the instantaneous intensity vector and diffuseness of the direct sound and early reflections derived from the B-Format RIRs, and by routing the non-diffuse components in the direction of the corresponding intensity vector using Vector Base Amplitude Panning (VBAP) rendering [13] in order to give a better "*hall sound*" impression in auralization of concert halls. The diffuse part is reproduced on all loudspeakers using Gaussian-shaped noises.

2. ANALYSIS OF MEASURED B-FORMAT ROOM IMPULSE RESPONSES

2.1 Intensity Vector and Diffuseness from B-Format Ambisonics

The Ambisonics approach is based on the solution of the wave equation in spherical coordinates. In any point in space, the acoustic pressure can be expressed by a Fourier-Bessel decomposition, where directional functions Y_{mn}^{σ} called spherical harmonics appear. These functions are associated with the weighting coefficients B_{mn}^{σ}.

$$p(kr,\theta,\delta)=\sum_{m=0}^{\infty} i^m\, j_m(kr)\sum_{n=0}^{m}\sum_{\sigma=\pm1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\delta)$$

where *k* represents the wave number, *r* the observed radius, θ and δ the azimuth and elevation angle respectively. The Fourier-Bessel decomposition must be truncated at a finite order *M* due to practical limitations. The accuracy of the reproduction and the size of the reconstructed sound field (listening area) depend on the order of the spherical harmonic functions. Hence, the sound field is described from a limited number of coefficients B_{mn}^{σ} (*m = 0, 1, …, M*) also called Ambisonics components. In the particular case of a plane wave of amplitude *S* coming from the direction (θ_S,δ_S), these components are defined by [4]:

$$B_{mn}^{\sigma} = Y_{mn}^{\sigma}(\theta_S,\delta_S)S$$

The equation describes the encoding process for a single sound source. Thus, the sound field is decomposed in the spherical harmonics Y_{mn}^{σ} evaluated at the direction of the source and multiplied by the wave amplitude S. The number of components *K* for a 3D Ambisonics system is calculated from the order *M*:

$$K=(M+1)^2$$

It follows that, for *M=1* there are four Ambisonics components. M Gerzon developed an encoding system for the first order Ambisonics called B-format and associated decoding methods [4]. In B-Format, the sound field is encoded by the first four Ambisonics components known as channels W, X, Y and Z. Channel W reflects the sound pressure component and the three following channels define its gradient, which are proportional to the particle velocity components. The first order Ambisonics SoundField microphone was built in 1977 [4,6]. It contains four sub-cardioid capsules set in a regular tetrahedron. B-Format channels are obtained by combining the capsules signals. Consequently, each B-format RIR is composed of four impulse responses.

The advantage of B-Format is that encoding and decoding steps are separated. In basic B-Format decoding, loudspeakers are generally considered to be regularly distributed on the reproduction area and all of them are always contributing jointly to the resynthesized sound field. A basic decoding process consists of projecting the encoded components on the spherical harmonic functions sampled at each loudspeaker direction. This mathematical decoding process is exact for a centered position but as frequency is increased the listening area for an accurate reproduction gets smaller. For the first order, 700Hz is the theoretical frequency limit in an area comparable to the size of an average head [4]. As a consequence the spatial image is perceptually blurred or unstable. In contrast, parametric decoding proposes to extract the main instantaneous directional information contained in the B-Format encoding. This information can then be exploited in the rendering system using various panning methods such as VBAP for instance [13]. Even though the parametric decoding results from an approximation of the sound field (e.g. direction of arrival and diffuseness or decomposition on two plane waves) it can give rise to a perceptually stable reproduction.

As was proposed in [7] and [14], in B-Format encoding, the acoustic pressure can be derived from W channel and the particle velocity vector from X, Y and Z channels. Knowing that the instantaneous energy density *E* and intensity *I* of a general acoustic field can be expressed in terms of the particle velocity vector ***v*** and the acoustic pressure *p* as [15]:

$$E(t) = \frac{1}{2}\rho[v^2(t) + z^{-2}p^2(t)]$$

$$I(t) = p(t)v(t)$$

where ρ, $z=\rho c$ and *c* represent the density, the impedance of the medium and the speed of sound respectively. The instantaneous intensity vector can be expressed in magnitude and direction as:

$$|I| = \frac{\sqrt{(WX)^2 + (WY)^2 + (WZ)^2}}{\rho c}$$

$$\theta = \tan^{-1}\left(\frac{Y}{X}\right)$$

$$\delta = \tan^{-1}\left(\frac{Z}{\sqrt{X^2 + Y^2}}\right)$$

where θ and δ represent the azimuth and elevation angle respectively. In the same way, the instantaneous energy density can be expressed as:

$$E = \frac{W^2 + X^2 + Y^2 + Z^2}{2\rho c^2}$$

Additionally, diffuseness of sound is calculated from the ratio of active intensity to energy defined by:

$$\psi = 1 - \frac{|I|}{Ec}$$

$$\psi = 1 - \frac{2\sqrt{(WX)^2 + (WY)^2 + (WZ)^2}}{W^2 + X^2 + Y^2 + Z^2}$$

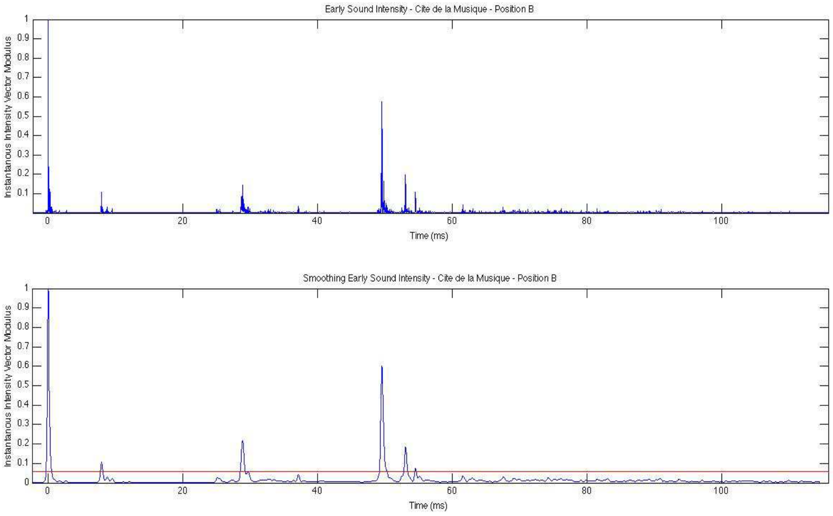
As diffuseness approaches zero, the net flow of energy comes from a single direction. As the value approaches one, it indicates a more diffuse sound. For calculations, it is important to note that, in SoundField microphones the X, Y and Z level channels are enhanced by 3dB in comparison with the W channel level.

2.2. Direct Sound and Early Reflections Processing

The processing of B-Format RIR is divided in early and late reflections. A reasonable approximation for the transition time between early reflections and late reverberation is defined by [16]:

$$t_{mixing} = \sqrt{V}$$

where t_{mixing} is the mixing time, expressed in *ms* and *V* is the volume of the room in m^3. Instantaneous intensity vector and diffuseness is calculated at each sample until the mixing time is reached. The process of identifying direct sound and main early reflections is realized in three steps (Figure 1). Firstly, the intensity vector modulus is smoothed and the diffuse average level around the mixing time is calculated. Secondly, only the samples with an intensity modulus 10 dB above the diffuse level are retained. Finally, the second derivative is calculated from the smoothed intensity signal formed by the retained samples in order to obtain peak indexes. In this way, direct sound information (magnitude and direction) is extracted for the first important peak (direct sound) and for the main reflections.



Delay [ms]	Level [dB]	Azimuth [°]	Elevation [°]
0	0,0	0,8	2,1
8	-22,4	6,8	-3,8
29	-19,3	-87,5	-45,8
30	-34,6	-61,3	-37,3
50	-5,5	0,8	-3,8
53	-16,2	5,5	-1,4
55	-22,4	1,5	6,4

Figure 1. Intensity vector modulus and smoothed intensity before mixing time for an individual B-Format RIR of *Cité de la Musique* concert hall. The red line indicates the threshold level (10 dB above the diffuse level). The table shows the direction of arrival for direct sound and main early reflections.

3. CONCERT HALL AURALIZATION

Auralization is made in the Institute’s listening room. It is a semi-anechoic room built on a floating floor with a reverberation time lower than 0.06 s for frequencies above 250 Hz and 0.25 s below. The reproduction system contains a subwoofer JBL 4645C and twelve loudspeakers Studer-A1, six forming a hexagon at ear’s level, three near the ceiling forming an equilateral triangle and three over the floor forming another equilateral triangle in opposite orientation. Acoustically transparent fabric panels hide the loudspeakers.

A subset of B-Format RIR database was selected covering different types of halls. It corresponds to measurements made in a central position for the source and the microphone. The halls selected were: Théâtre de l'Athénée, Bastille Opera House, Théâtre du Châtelet, Cité de la Musique, Salle Cortot, Garnier Opera House, Louvre Auditorium, and Salle Pleyel.

Direct sound and main early reflections are divided in diffuse and non-diffuse part by using the diffuseness factor. For the non-diffuse part, auralization is made through the VBAP method. Thus, sound is rendered by a maximum of three loudspeakers using the intensity magnitude and direction information. The diffuse part is reproduced on all loudspeakers using Gaussian-shaped noises different for each loudspeaker. For non-main first reflections, that is, those below threshold (see Sect. 2.2), and for late reverberation, *basic* Ambisionics decoding is retained. In this way, for each B-Format RIR, one impulse response is calculated for each loudspeaker in the listening room according to its coordinates.

In order to sharpen the direct sound localization, a thirteenth loudspeaker was installed in front of the listener's position at zero azimuth and elevation position. B-Format sound field rotation is made to reproduce the direct sound just from this loudspeaker. Auralization is then obtained by convolving with an anechoic signal the thirteen impulse responses, one for each loudspeaker, previously mentioned. To compensate the non-perfectly regular placement of the loudspeakers in the room, the gain and delay adjustments were made in listener's position. Furthermore, the whole system was equalized according to the frequency response of the thirteenth loudspeaker because the loudspeakers frequency responses were well comparable. The signal processing hardware is composed of a DIGI96 soundcard and two RME ADI-8 Pro converters. The auralization application was developed in MAX/MSP exploiting HISS tools [17] to enable multi-channel convolution in real time.

4. ANALYSES AND RESULTS

The auralization method is evaluated in two ways. First, by calculating the acoustical descriptors of the convolved sound field and by comparing them with the reference sound field (*in-situ* RIR measurements) using Just Noticeable Difference (JND) criteria. Secondly, by plotting the instantaneous intensity vector around the direct sound and main early reflections for reference sound field and by comparing them with the convolved sound field, using first-order Ambisonics and VBAP rendering.

The same excitation signal as for *in-situ* measurements was used. A SoundField ST250 microphone was placed at the listener's position for measuring the convolved sound field. The six conventional acoustic indices - reverberation time (T30), early decay time (EDT), clarity (C80), the central time (Ts), the sound amplification (G) and the lateral factor (LFC) - were analyzed in octave bands from 125Hz to 4000Hz. All indices were calculated from the omnidirectional impulse response related to the W component and from the bidirectional left-right impulse response related to the Y component for LFC. In order to evaluate if reference and convolved acoustic parameters give the same perceptual impression, the six indices were averaged over several octave bands according to Annex A in ISO 3382-1:2009 standard [18]. Table 1 shows the reference average values for each hall. G Factor value was taken from [19].

HALL	G [dB]	C80 [dB]	Ts [ms]	EDT [s]	RT30 [s]	LFC	V [m^3]	Dist [m]
Athénée	7,8	3,9	68	1,05	1,05	0,23	3366	8,6
Bastille	2,4	3,1	85	1,73	1,67	0,23	26000	19,3
Châtelet	0,2	2,1	83	1,34	1,51	0,24	8900	12
Cite	4,35	-2,5	148	1,76	1,84	0,22	13400	17,9
Cortot	10	3,8	75	1,09	1,25	0,16	3400	6,3
Garnier	4,8	2,1	79	1,47	1,22	0,15	10000	14,3
Louvre	8,7	2,3	93	1,26	1,50	0,14	4500	7,7
Pleyel	6	1,6	98	1,74	1,77	0,14	17800	8,3

Table 1. Average acoustic indices, volume and measurement distance for each hall.

Table 2 shows the differences between reference and convolved sound field for each acoustic index in terms of JND. All differences are within 1 JND except for LFC in *Cité de la Musique*. This indicates that the reference sound field is perceptually comparable to the convolved sound field.

To analyze the improvement in sound incidence reproduction, the instantaneous intensity vector for direct sound and main early reflections is plotted for the three sound fields (reference, first-order Ambisonics decoding and VBAP rendering) in a window of 500µs centered on the main peak. For the halls analyzed, the graphics point out that with the VBAP approach, the direct sound and the main early reflections of the convolved sound field are more similar to the reference than with the *basic* Ambisonics approach.

HALL	G	C80	Ts	EDT	RT30	LFC
Athénée	0,2	0,3	0,5	0,5	0,1	0,6
Bastille	0,4	0,6	0,8	0,0	0,0	0,4
Chatelet	0,3	0,3	0,3	0,9	0,1	0,8
Cite	0,2	0,7	0,4	0,0	0,1	1,2
Cortot	0,3	0,1	0,1	0,4	0,1	0,4
Garnier	0,1	0,5	0,4	0,6	0,2	0,9
Louvre	0,1	0,0	0,2	0,0	0,4	0,5
Pleyel	0,4	0,5	0,5	0,4	0,2	0,7

Table 2. Differences between reference and the convolved sound field in terms of JND

Figure 2 shows two examples, one for direct sound and the other for one frontal reflection. The blue lines indicate the reference sound and the red lines the convoluted sound using VBAP or *basic* first-order Ambisonics rendering. The four graphics in the left belong to Bastille Opera House for direct sound. Ambisonics reproduction shows a greater variation in the direction of sound, which could bias the localization towards the nearest loudspeaker in the indicated direction, making direct sound fuzzy. On the other hand, VBAP gives a narrow and accurate reproduction in the direction of the reference sound, which could provide a clear and easy direct sound localization. An improvement is also observed for the main early reflections, as visible in the right of Figure 2 for one reflection. In addition, informal listening tests showed a better "hall sound" impression than with *basic* first order Ambisonics reproduction.

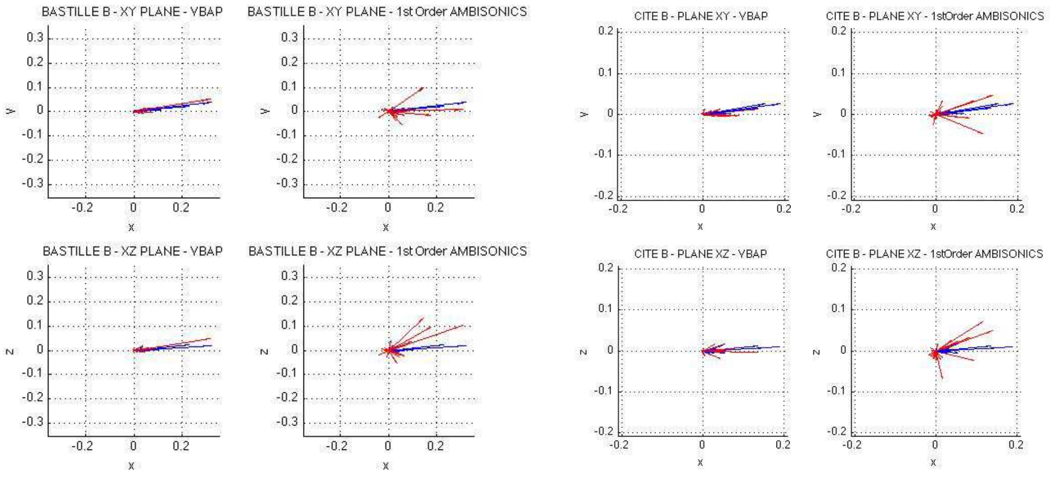


Figure 2. Bastille Opera House direct sound intensity vector (left). Cite de la Musique front reflection intensity vector (right). Blue lines indicate the reference sound and red lines the convoluted sound using VBAP or *basic* first-order Ambisonics rendering.

5. CONCLUSION

A method was presented to improve the sound incidence reproduction from B-Format room impulse responses. The level and direction of arrival information of the direct sound and the main early reflections are extracted from DRIRs in order to be rendered by the VBAP method. Late reflections and non-main early reflections are reproduced with a basis B-Format decoder. The convolved sound field showed to be comparable to the reference sound field using JND criteria for the classic acoustics indices. Furthermore, VPAB presents a more narrow and accurate direct sound and early reflections reproduction than Ambisonics. Formal listening tests will be conducted to assess the improvement in direct sound localization and spatial impression.

6. ACKNOWLEDGEMENTS

This work is part of a doctorate thesis supported by Los Andes University in Bogotá, Colombia.

7. REFERENCES

[1] M. Kleiner, B-I. Dalenback, P. Svensson. “Auralization – an overview”. J. Audio Eng. Soc., Volume 41, Number 11, pp. 861-875, November 1993.

[2] T. Lokki. “Recording and reproducing concert hall acoustics for Subjective evaluation”. International Seminar on Virtual Acoustics (ISVA). Valencia, November 2011.

[3] H. Møller. “Fundamentals of binaural technology”. Applied Acoustics, Volume 36, Number 3-4, pp. 171–218, 1992.

[4] J. Daniel. “Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia”. PhD Thesis, Université Pierre et Marie Curie, Septembre 2000.

[5] C. Guastavino. “Etude sémantique et acoustique de la perception des basses fréquences dans l’environnement urbain”. PhD Thesis, Université Pierre et Marie Curie, 2003.

[6] S. Bertet, J. Daniel, E. Parizet, O. Warusfel. “Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources”. Acta Acustica united with Acustica, Volume 99, Number 4, July/August 2013.

[7] J. Merimaa and V. Pulkki. “Spatial impulse response rendering I: Analysis and synthesis”. J. Audio Eng. Soc., Volume 53, Number 12, pp. 1115–1127, December 2005.

[8] S. Berge, N. Barrett. “High angular resolution planewave expansion”. In Proc. of the 2nd International Symposium on Ambisonics and Spherical, Paris, France, May 2010.

[9] J. Merimaa and V. Pulkki. “Spatial impulse response rendering”. In Proc. of the 7th International Conference on Digital Audio Effects, Naples, Italy, October 2004.

[10] J-D. Polack and F. Leão Figueiredo. « Room acoustics auralization with Ambisonics ». In IOA annual meeting, Nantes, April 2012.

[11] M. Barron and A.H. Marshall. “Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure”. Journal of Sound and Vibration, Volume 77, Issue 2, pp. 211-232, July 1981.

[12] D.H. Griesinger,”The Relationship Between Audience Engagement and Our Ability to Perceive the Pitch, Timbre, Azimuth and Envelopment of Multiple Sources" A PowerPoint presentation given to AES local sections in Boston and Washington DC, June 2010.

[13] V. Pulkki. « Virtual sound source positioning using vector base amplitude panning ». J. Audio Eng. Soc., Volume 45, Number 6, pp. 456–466, June 1997.

[14] A. Farina, E. Ugolotti. “Subjective comparison between stereo dipole and 3d Ambisonic surround systems for automotive applications”. In Proc. AES 16th International conference on Spatial Sound Reproduction, April 1999.

[15] G. Schiffrer, D. Stanzial. “Energetic properties of acoustic fields”. J. Acoust. Soc. Am. Volume 96, Issue 6, pp. 3645-3653. 1994.

[16] G. Defrance, J-D. Polack. “Measuring the mixing time in auditoria”. Acoustics’08, Paris. June-July 2008.

[17] HISS Website. Available: http://www.thehiss.org.

[18] ISO3382-1:2009. ‘‘Measurement of room acoustic parameters. Part 1: Performance Spaces”.

[19] F.L. Figueiredo. “Indices acoustiques et leurs rapports statistiques : vérification objective et subjective pour un ensemble de salles de spectacles”. PhD Thesis. Université Pierre et Marie Curie. 2011.