Objective comparison between Ambisonics basic decoding and a SIRR-based parametric decoding in the context of concert hall auralization

Juan Pablo ESPITIA HURTADO[1,2]; Jean-Dominique POLACK[1,2]; Olivier WARUSFEL[3]

[1] UPMC Univ Paris 06, UMR 7190, Institut Jean Le Rond d'Alembert, F-75005 Paris, France

[2] CNRS, UMR 7190, Institut Jean Le Rond d'Alembert, F-75005 Paris, France

[3] UMR 9912, Sciences et Techniques de la Musique et du Son, IRCAM-CNRS-UPMC, F-75004 Paris, France

ABSTRACT

The room acoustics group at Pierre and Marie Curie University has a database of directional room impulses responses (DRIR) measured in unoccupied concert halls and theatres in Paris. The DRIRs were measured in 2009 with a SoundField ST250 microphone in B-Format (first order Ambisonics) for auralization purposes. Listening tests conducted in 2012, using a basic Ambisonics reproduction over twelve loudspeakers, showed a lack of spaciousness that could be linked to a high interaural coherence and a non-optimal sound incidence reproduction. Decoding improvement is made by means of the estimation of the energy and the intensity vector of the sound field, based on Spatial Impulse Response Rendering (SIRR) method. The constant Q transform (CQT) is used for time and frequency domain analysis. The non-diffuse components are routed using VBAP rendering and diffuse field is synthesized using MLS signals. The intensity vector associated to the direct sound, the reverberation and the interaural correlation profile are compared between this decoder and a basic Ambisonics decoder. Finally, a comparison of some conventional acoustic descriptors is performed between the real and reproduction contexts.

Keywords: Room Acoustics, Auralization I-INCE Classification of Subjects Number(s): 74.9,76.9

1. INTRODUCTION

Auralization is a useful and widely used tool for subjective evaluation of concert halls. Kleiner (1) defined auralization as the process of rendering audible the sound field of a source in a space. Contrary to in-situ listening tests, auralization allows comparison between different spaces with exactly the same musical source in the same listening conditions. Furthermore, comparisons can be done without time lapses which enables listening carefully to the differences in acoustics between concert halls. However, the relevance of the results depends on the degree of fidelity of the virtual rendition to the real auditory environment. In general terms, concert hall 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 by the encoding and decoding process of the 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 is not suitable as the rendition will be marred by 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 Head Related Transfer Function, HRTF (2). 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. In this respect, the first-order Ambisonics B-Format or its High Order Ambisonics (HOA) extensions are good candidates (3). 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 localisation accuracy (4). As an alternative, HOA increases angular discrimination and enlarges the available listening

[1] espitia@lam.jussieu.fr

area (the higher the order, the better the spatial resolution (5)). But HOA requires high spatial resolution microphones (e.g. spherical microphone arrays') for measuring DRIR, as well as a large number of loudspeakers for decoding. Recently, other methods have been proposed to exploit B-Format DRIRs using parametric decoding. This is the case for Spatial Impulse Response Rendering technique (SIRR) (6), employing sound intensity theory and High Angular Resolution Planewave Expansion (HARPEX) (7), based on plane wave decomposition. In both cases, B-Format signals are analysed in time and frequency, in order to improve the spatial image of sound. Listening tests for both methods were compared with first-order Ambisonics systems showing better results (7) (8). The SIRR technique has been widely used in concert hall evaluation by the Virtual Acoustics research team at the department of Media Technology at the Aalto University School of Science.

The room acoustics group at the d'Alembert Institute at the Université Pierre et Marie Curie has a data base of B-Format RIRs as measured from 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 15 seconds exponential sweep-sine from 20 Hz up to 20 kHz was used as the 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 (centre, left and right). Furthermore, between 2010 and 2012, listening tests were also conducted from those measurements using a first-order Ambisonics basic decoder in a listening room consisting of 12 loudspeakers positioned in dodecahedral form (9). Results showed, amongst other weaknesses, a lack of spaciousness that could be linked to non-optimal sound incidence reproduction. This paper studies the improvement of spatial rendering achieved by exploiting sound intensity theory for decoding B-format RIRs. The merit of the method is estimated through the comparison of the intensity vector associated to the direct sound, of the reverberation profile, of the interaural cross-correlation profile and of some conventional acoustical descriptors between real and reproduction contexts. The decoding method is based on SIRR (6), although the time frequency processing is made using Constant Q Transform and the diffuse field is rendered using modified reciprocal Maximum Length Sequence (MLS) signals (10). Also, one centrally-positioned speaker is dedicated to emit only non-diffuse signal (i.e. direct sound and frontal first reflections). The reason for this, is taken from Griesinger (11) who suggests that if direct sound is clearly distinct, as is the case with accurate localisation, it is possible for the brain to separate this perception from the perception of reflections and reverberation and in consequence to better perceive enveloping sound. Thus, decoding improvement was achieved by estimating the instantaneous intensity vector and diffuseness 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 (12), and the diffuse part is reproduced on all loudspeakers using modified reciprocal MLS signals, in order to give a better 'hall sound' impression in auralization of concert halls.

2. DIRECTIONAL ROOM IMPULSE RESPONSES

2.1 B-Format first order Ambisonics

The Ambisonics approach (3) 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}.

$$p(kr,\theta,\delta) = \sum_{m=0}^{\infty} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\delta) \tag{1}$$

where k represents the wave number, r the observed radius, θ and δ the azimuth and elevation angles 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 by a limited number of coefficients B_{mn}^{σ} $(m = 0,1,\ldots,M)$ 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

$$B_{mn}^{\sigma} = Y_{mn}^{\sigma}(\theta_S,\delta_S)S \tag{2}$$

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:

$$K = (M+1)^2 \tag{3}$$

It follows that, for M=1 there are four Ambisonics components. M. Gerzon developed an encoding system for first order Ambisonics called B-format and associated decoding methods (3). 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 (3) (5). 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 over the reproduction area and all of them are always contributing jointly to the re-synthesized sound field. A basic decoding process consists of projecting the encoded components on the spherical harmonic functions sampled at each loudspeaker position. This mathematical decoding process is exact for a central position but as frequency is increased the listening area for accurate reproduction gets smaller. For the first order, 700 Hz is the theoretical frequency limit in an area comparable to the circumference of an average head (3). As a consequence the spatial image is perceptually blurred or unstable. In contrast, parametric decoding (e.g. SIRR) 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. 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 perceptually stable reproduction.

2.2 Intensity Vector and Diffuseness

SIRR decoding is based on the directional analysis and the estimation of the diffuseness of the sound field. As is well known, the instantaneous energy density E and the instantaneous sound intensity $\mathbf{I}$ of a general acoustic field can be expressed in terms of the particle velocity vector $\mathbf{u}$ and the acoustic pressure p as:

$$E = \frac{1}{2}\rho_0(Z_0^{-2}p^2 + \mathbf{u}^2) \tag{4}$$

$$\mathbf{I} = p\mathbf{u} \tag{5}$$

where ρ_0, $Z = \rho_0 c$ and c represent the density, the impedance of the medium and the speed of sound respectively. Thus, the vector $\mathbf{I}$ expresses the magnitude and direction of the instantaneous flow of sound energy per unit area. Additionally, in energetic analysis, *diffuseness estimate* ψ is defined as the proportion of the active intensity to the energy density (8).

$$\psi = 1 - \frac{\|\langle \mathbf{I} \rangle\|}{c\langle E \rangle} = 1 - \frac{2Z_0\|\langle p\mathbf{u} \rangle\|}{\langle p^2 \rangle + Z_0^2\langle \mathbf{u}^2 \rangle} \tag{6}$$

where $\langle \cdot \rangle$ represents the expectation operator, and $\| \cdot \|$, the ℓ_2 norm. In an ideally diffuse sound field the ψ value approaches one. As ψ approaches zero, the net flow of energy comes from a single direction.

2.3 Constant-Q Transform

Instead of using short-time Fourier transform (STFT), as originally implemented in SIRR (6), our method uses Constant Q Transform (CQT) for time-frequency processing of the DRIRs. CQT is a technique that transforms a time domain signal into the time-frequency domain so that the centre frequencies of the frequency bins are geometrically spaced, their Q-factors all being equal (13). Thus, CQT gives a better trade-off between temporal and spectral resolution for musical signal analysis than STFT with regard to the human hearing response. That is, the spectral resolution is better for low frequencies whilst temporal resolution is better in high frequencies which in our case is favorable for the analysis, the synthesis and the visualization of the DRIRs for auralization purposes.

Given the signal $x(n)$, the CQT representation $X^{CQ}(k,n)$, is defined as (14)

$$X^{CQ}(k,n) = \sum_{m=0}^{N} x(m)a_k^*(m-n) \tag{7}$$

where k is the frequency bin, n refers to the time frame, N is the length of the signal $x(n)$ and the time-frequency atoms $a_k^*(m)$ are the complex conjugated functions defined by

$$a_k(m) = g_k(m)\exp(i2\pi m\frac{f_k}{f_s}) \tag{8}$$

where f_k is the centre frequency of bin k, f_s the sampling rate and $g_k(m)$ is the window function. The centre frequencies f_k are computed as

$$f_k = f_0 2^{\frac{k}{b}} \tag{9}$$

where f_0 represents the centre frequency of the lowest-frequency bin, and b is the number of bins per octave.

The CQT method was firstly introduced by Brown and co-workers (15). However, CQT was not widely used in music signal analysis due to the lack of an inverse transform for a perfect reconstruction of the original signal and to the complexity of the data structure. In 2010, dealing with these drawbacks Schöerkhuber and Klapuri (13) developed a computationally efficient toolbox allowing an acceptable reconstruction of the signal. More recently, Schörkhuber and co-workers (14) improved the computation of CQT giving an efficient framework that allows a perfect reconstruction. Additionally, the time resolution in low frequencies can be improved by decreasing the Q-factors of the low frequencies bins. Thus, the time-frequency processing of the DRIRs in this work are made using CQT representation with the help of the MATLAB toolbox described in (14).

3. SIRR-BASED DECODING

3.1 Analysis of measured B-Format RIR

As was proposed in (6), in B-Format encoding, the acoustic pressure p can be derived from channel W and the particle velocity vector $\mathbf{u}$ from channels X, Y and Z. Thus, defining $W(n,k)$, $X(n,k)$, $Y(n,k)$ and $Z(n,k)$ the CQT representation of the B-format signals $w(n)$, $x(n)$, $y(n)$ and $z(n)$ respectively, the acoustic pressure $P(k,n)$ and the particle velocity vector $\mathbf{U}(k,n)$ can be computed as

$$P(k,n) = W(k,n) \tag{10}$$

$$\mathbf{U}(k,n) = \Big[X(k,n), Y(k,n), Z(k,n)\Big]^T \tag{11}$$

where k is the frequency bin and n refers to the time frame. From equations 4, 5 and 6 the energy density $E(n,k)$, the active intensity vector $\mathbf{I_a}(k,n)$ and the diffuseness estimator $\psi(k,n)$ is calculated as:

$$E(k,n) = \frac{1}{2\rho_0 c^2}(|W(k,n)|^2 + \frac{1}{2}\|\mathbf{U}(k,n)\|^2) \tag{12}$$

$$\mathbf{I_a}(k,n) = \frac{1}{\sqrt{2}\rho_0 c}\,\mathrm{Re}\{W^*(k,n)\mathbf{U}(k,n)\} \tag{13}$$

$$\psi(k,n) = 1 - \frac{\sqrt{2}\|\langle \mathrm{Re}\{W^*(k,n)\mathbf{U}(k,n)\}\rangle\|}{\langle |W(k,n)|^2 + \frac{1}{2}\|\mathbf{U}(k,n)\|^2\rangle} \tag{14}$$

where $*$ denotes the complex conjugated and $\mathrm{Re}\{\}$ the real operator. It is important to note that the preceding equations take into account that in SoundField microphones the levels of X, Y and Z channels are enhanced by 3dB compared to the level of the W channel.

The CQT processing is established by thirds of an octave from 42 Hz to 22050 Hz ($b = 3$ and $f_0 = 42$ Hz in equation 9). Additionally, Q factor is decreased in low frequency range to improve time resolution in low frequencies. The final time resolution goes from 20 ms for the lowest frequency band to 0.125 ms in the highest frequency band.

For each frequency-time frame the magnitude and direction of the instantaneous intensity vector is calculated. In the same way, as indicated in the Equation 6, diffuseness is estimated with the help of a moving average filter along all time frames for each frequency bin. It was observed that the window size of the filter

had an important effect on the estimated diffuseness, as previously noted by DeGaldo and co-workers (16), and that it depends on the central frequency which is also related to the CQT processing. On the other hand, long window size in the early part of the impulse response, can lead to overestimating the diffuseness factor and then, in synthesis process, to smoothing the directional character of the direct sound and "specular" reflections. Hence, in order to avoid this kind of issue, the window size was fixed at 1 ms for all frequency bins in the early part of the impulse response. In the late part the window size was fixed according to the central frequency, being of decreasing length the higher the frequency.

3.2 DRIR Synthesis

Synthesis is made according to the loudspeakers coordinates (cf. 4) in the listening room of the Institute. In this way, for each B-Format RIR, one impulse response is calculated for each loudspeaker. The synthesis process is divided into non-diffuse and diffuse parts using the diffuseness estimator $\psi(k,n)$ at each frequency-time frame, by multiplying the w channel signal by $\sqrt{1-\psi(k,n)}$ in the first case and by $\sqrt{\psi(k,n)}$ in the second case.

On the one hand, the non-diffuse part, is rendered by a maximum of three loudspeakers through the VBAP method using the directional information of the instantaneous intensity vector. On the other hand, the diffuse part is rendered on all loudspeakers using reciprocal Maximum Length Sequence (MLS) signals for each loudspeaker. As is well known, low values in IACC correspond to a high degree of spaciousness in the perception of enveloping reverberance. As mentioned in (17), MLS-pairs have low values of cross-correlation, hardly found in other random noise signals. These signals have been used by Xiang and co-workers (17) in controlling and synthesising the reverberant part of binaural room impulse responses.

Thus, twelve different MLS signals are generated and processed using CQT representation. The magnitude of each frequency-time frame for each independent MLS signal is equalized to the magnitude of each frequency-time frame of the W-channel $W(n,k)$ signal. Thus, only phases are made random. Further, an additional gaining adjustment should be applied to warrant the same energy as in W-channel in each frequency bin when all MLS signals are superposed. Finally, for each loudspeaker, the inverse-CQT is applied independently to the non-diffuse part and to the diffuse part. Consequently, the final impulse response for each loudspeaker is the summation of the two parts (non-diffuse and diffuse sound) in time-domain.

In some cases, and for some frequencies, it was necessary to use only VBAP to obtain the values of some acoustic indices within 1 JND (Just Noticeable Difference) by comparison to the reference sound field (cf. 5) but that matter is not explained in detail in this paper.

As mentioned previously in (18) the polar patterns and frequency responses of each channel of the Sound Field ST250 microphone were measured in the anechoical room of LNE (Laboratoire National de Métrologie et d'Essais). Results showed that the polar directivity follows the theoretical curves between 125 Hz and 2 kHz in three dimensions and between 125 Hz and 4 kHz in the horizontal plane. Is important to note that both the diffuseness and intensity vector are calculated according to the Institute's microphone characteristics.

4. AURALIZATION

Auralization is achieved in the Institute's listening room. It is a small semi-anechoic room (2.77 x 3.24 x 3.62 m) 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 des Abesses (ABE), Théâtre de l'Athénée (ATH), Bastille Opera House (BAS), Théâtre du Châtelet (CHA), Cité de la Musique Concert Hall (CIT), Salle Cortot (COR), Garnier Opera House (GAR), Louvre Museum Auditorium (LOU), Orsay Museum Auditorium (ORS) and Salle Pleyel (PLE).

Table 1 – Concert hall volumes and measurement distances

	ABE	ATH	BAS	CHT	CIT	COR	GAR	LOU	ORS	PLE
Volume, m^3	1800	3366	26000	8900	13400	3400	10000	4500	1700	17800
Distance, m	6.25	8.6	19.3	12	17.9	6.3	14.3	7.7	13.2	8.3

In order to sharpen direct sound localisation, a thirteenth loudspeaker was installed in front of the listener's position at zero azimuth and elevation position (best position). In addition, B-Format sound field rotation is

made to reproduce the direct sound from this loudspeaker. Auralization is then obtained by convolving an anechoic signal with the thirteen impulse responses, one for each loudspeaker, as previously mentioned. To compensate for the imperfectly regular placement of the loudspeakers in the room, gain and delay adjustments were made in the listener position. Furthermore, because the loudspeakers frequency responses were well comparable, the whole system was equalized only according to the thirteenth loudspeaker. 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 (19) to enable multichannel convolution in real time.

5. ANALYSES AND RESULTS

The auralization method is evaluated in different ways. First, by plotting the instantaneous intensity vector around the direct sound for the reference sound field (DRIR in-situ measurement) and by comparing them with the convolved sound field, using first-order Ambisonics (FuMa weighting) and SIRR-based decoding. Secondly, through the comparison of reverberation profiles and inter-aural cross correlation profiles. Finally by calculating selected acoustical descriptors of the convolved sound field and by comparing them with the reference sound fields using Just Noticeable Difference (JND) criteria. The same excitation signal as for in-situ measurements was used. A SoundField ST250 microphone was placed at the center of the loudspeaker hemisphere for measuring the convolved sound field using either first-order Ambisonics and SIRR-based decoding. Both decoding systems were measured using the thirteen loudspeaker configuration (cf. 4).

5.1 Intensity Vector of Direct Sound

To assess the improvement in sound incidence reproduction, the instantaneous intensity vector for direct sound is plotted for the three sound fields (reference, first-order Ambisonics decoding and SIRR-based decoding) in a window of 1 ms centered on the main peak. For the halls analysed, the graphics point out that with the SIRR-based decoding, the direct sound of the convolved sound field is more similar to the reference sound field than to the first-order Ambisonics basic decoding.

As an example, the Figure 1 shows the instantaneous intensity vector in the horizontal and medial planes for the three sound fields at different CQT frequency-time frames, around 1 ms of direct sound, between 125 Hz and 4 kHz. The calculations are taken from a DRIR of the Théâtre du Châtelet.

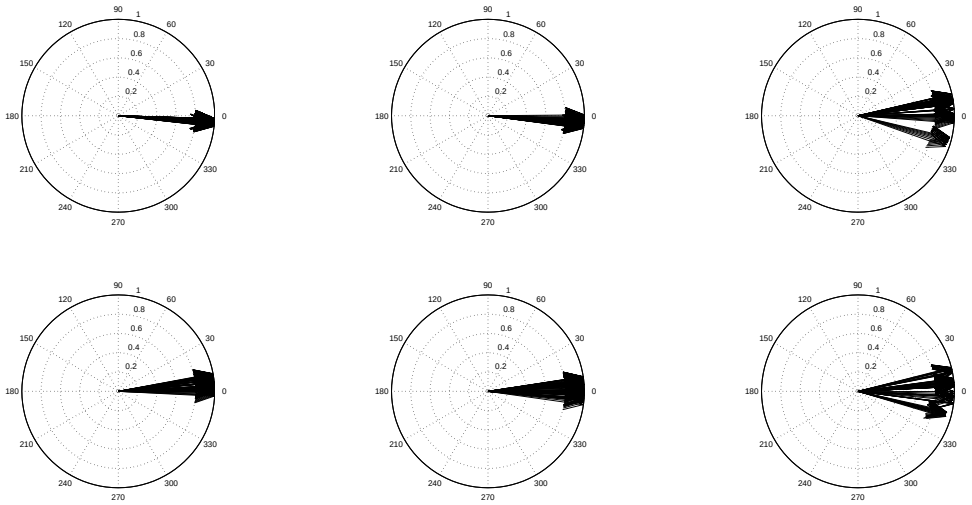

Figure 1 – Direction of the Intensity Vector for Direct Sound for three sound fields. Reference (left), SIRR-based (centre) and first-order Ambisonics (right). Horizontal Plane (above). Medial Plane (below)

.

As can be observed, Ambisonics reproduction shows greater variation in the direction of sound, which can make difficult the localisation of direct sound and can blur the image sound. On the other hand, SIRR-based decoding gives narrow and accurate reproduction in the direction of the reference sound, which could accurately distinguishes direct sound and its localisation.

5.2 Reverberation Profile

The reverberation time is compared for the three sound fields using RT30 between 63 Hz and 8 kHz for the 10 halls. It was observed that the reverberation profile in Ambisonics basic decoding and SIRR-based decoding are similar (e.g. below 1 JND for all frequencies) to the reverberation profile of the measured impulse response of the reference sound field (i.e. W-channel). Figure 2 shows the RT30 of two halls for the three sound fields.

We can conclude that either first-order Ambisonics decoding and SIRR-based decoding are robust in recreating the sound level decay of the reference sound field.

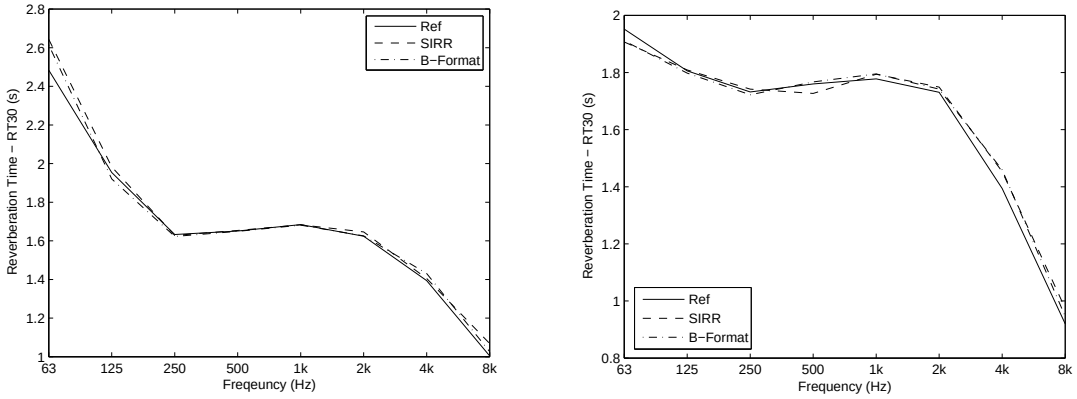

Figure 2 – Reverberation profiles of Bastille Opera House (left) and Salle Pleyel (right) for three sound fields (Reference, SIRR-based and first-order B-Format Ambisonics)

5.3 Interaural Cross-Correlation Profile

Improvement in the reproduction of the diffuse part is assessed by the comparison of the two methods' Interaural Cross-Correlation Coefficient (IACC). No binaural measurement was made either in-situ or in reproduction context. Instead, the different loudspeakers' impulse responses, calculated both in first-order Ambisonics (SN3D and FuMa weightings) and SIRR-based decoding, were directly convolved from a Head Related Impulse Response (HRIR) database according to the loudspeakers' coordinates in the Institute' listening room. Consequently, the binaural room impulse responses (BRIRs) were obtained from the reproduced sound fields (first-order Ambisonics and SIRR-based decoding).

The IACC profile is calculated from Equations 15 and 16 on the late part of the BRIR (t_1= 80 ms, t_2= ∞) in the octave bands between 63 Hz and 8 kHz.

$$IACC_{t_1,t_2} = \max_{\tau} |IACF_{t_1,t_2}(\tau)|, \tau \in (-1,1)ms \tag{15}$$

$IACF_{t_1,t_2}(\tau)$ denotes the Interaural Cross-correlation Function defined as

$$IACF_{t_1,t_2}(\tau) = \frac{\int_{t_1}^{t_2} p_l(t)p_r(t+\tau)\mathrm{d}t}{\sqrt{\int_{t_1}^{t_2} p_l^2(t)\mathrm{d}t \int_{t_1}^{t_2} p_r^2(t)\mathrm{d}t}} \tag{16}$$

where $p_l(t)$ and $p_r(t)$ are the impulse response at the entrance to the left and right ear canals respectively.

The Figure 3 shows the IACC profiles calculated from the DRIR of both Bastille Opera House and Cité de la Musique Concert Hall. As can be seen, compared to first-order Ambisonics decoding using SN3D weighting, SIRR-based decoding shows lower IACC values for frequencies above 500 Hz. However, when compared to first-order Ambisonics decoding using FuMa weighting, SIRR-based decoding presents a similar values up to 2kHz, but presents lower IACC values above this frequency.

5.4 Conventional Acoustic Index

Five conventional acoustic indices - early decay time (EDT), clarity (C80), the central time (Ts), the sound amplification (G) and the lateral factor (LFC) - were analysed in octave bands from 125Hz to 4 kHz

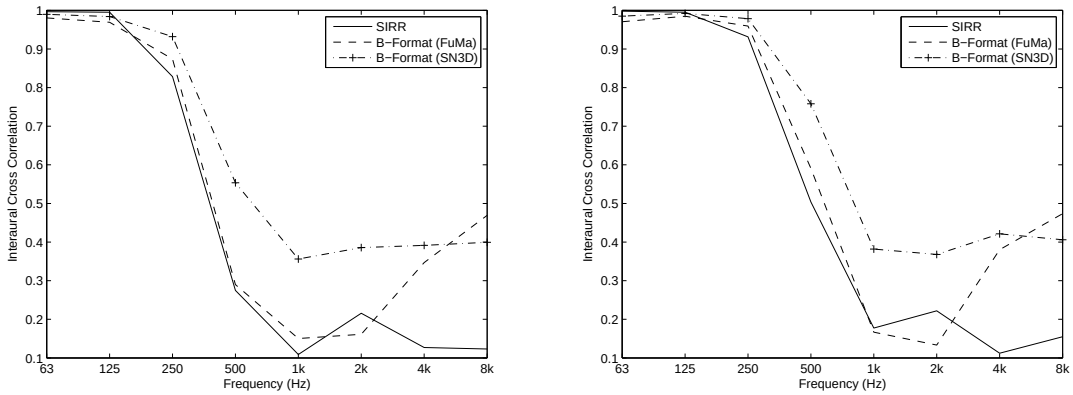

Figure 3 – Inter-aural Cross Correlation profile. Bastille Opera House (left) and Cité de la Musique Concert Hall (right). SIRR-based decoding (—). First-order B-Format Ambisonics decoding FuMa (- -) SN3D (-+)

as recommended in ISO 3382-1:2009 standard. 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. The Table 2 shows the values of the acoustic indices calculated from the in-situ measured DRIR (reference sound field) for the 10 halls selected. G Factor value was taken from (9).

In order to evaluate if reference and convolved acoustic parameters (first-order basic Ambisonics and SIRR-based decoding) give the same perceptual impression, the five indices were averaged in low (125 and 250 Hz), mid (500 and 1000 Hz) and high (2000 and 4000 Hz) frequencies and compared in terms of the Just Noticeable Difference (JND).

Table 2 – Acoustic Indices from In-Situ Measurements. L,M and H denote Low, Mid and High respectively

	ABE	ATH	BAS	CHT	CIT	COR	GAR	LOU	ORS	PLE
EDT-L, s	0.88	1.52	1.94	1.47	1.45	0.89	1.44	1.34	1.13	1.96
EDT-M, s	1.09	1.05	1.73	1.34	1.76	1.09	1.47	1.26	0.85	1.74
EDT-H, s	1.19	0.86	1.62	1.19	1.79	1.07	1.21	0.93	1.33	1.58
Ts-L, ms	84	98	145	139	141	79	129	123	98	116
Ts-M, ms	73	68	85	83	148	75	79	93	70	98
Ts-H, ms	70	59	88	87	140	69	71	65	104	86
C80-L, dB	4.8	3.2	-1.0	4.1	-2.7	3.5	-1.5	-0.7	1.6	0.8
C80-M, dB	3.1	3.9	3.1	2.1	-2.5	3.8	2.1	2.3	4.6	1.6
C80-H, dB	3.6	4.8	2.6	1.3	-1.9	3.4	3.2	4.5	0.9	2.4
G-L, dB	11.9	8.5	-0.2	-1.6	4.0	7.0	4.2	10.0	10.3	7.2
G-M, dB	9.3	7.8	2.4	0.2	4.4	10.0	4.8	8.7	10.7	6.0
G-H, dB	10.1	6.9	2.4	3.6	4.7	11.9	4.6	9.0	11.9	5.2
LFC-L	0.22	0.15	0.23	0.25	0.19	0.14	0.13	0.11	0.14	0.11
LFC-M	0.32	0.31	0.23	0.24	0.26	0.19	0.16	0.18	0.27	0.17
LFC-H	0.32	0.38	0.34	0.35	0.34	0.32	0.25	0.32	0.44	0.25

Table 3 shows the differences between reference and convolved sound fields (B-Format first-order Ambisonics and SIRR-based respectively) for each acoustic index in terms of JND. Results are reported as the mean and standard deviation (SD) of the JND values of the ten halls. Also, the minimum and maximum values are presented.

As can be seen, either with B-Format first-order Ambisonics and SIRR-based decoding for the indices EDT, Ts, G, C80 and LFC-L almost all differences are within 1 JND. Some high values are found in EDT-H, Ts-M and C80-M in B-Format basic decoding. Concerning LFC index, some significant differences are observed in both decoding systems at mid and high frequencies. However, in SIRR-based decoding the difference is less pronounced in high frequencies than in B-Format basic decoding.

Table 3 – Differences between reference and convolved sound fields in terms of JND

	B-Format basic decoding				SIRR-based decoding			
	Mean	SD	Min	Max	Mean	SD	Min	Max
EDT-L,	0.48	0.37	0.06	1.38	0.34	0.29	0.01	1.07
EDT-M,	0.52	0.31	0.19	1.21	0.44	0.36	0.01	1.11
EDT-H,	0.77	0.46	0.18	1.80	0.43	0.30	0.03	0.98
Ts-L,	0.29	0.28	0.01	0.87	0.45	0.34	0.06	1.10
Ts-M,	0.62	0.39	0.11	1.41	0.39	0.25	0.01	0.97
Ts-H,	0.34	0.18	0.05	0.65	0.29	0.19	0.03	0.59
C80-L,	0.31	0.29	0.02	0.90	0.61	0.37	0.09	1.17
C80-M,	0.65	0.47	0.04	1.52	0.49	0.37	0.01	1.03
C80-H,	0.27	0.15	0.06	0.54	0.34	0.18	0.04	0.59
G-L,	0.29	0.20	0.09	0.71	0.21	0.13	0.01	0.48
G-M,	0.27	0.14	0.01	0.49	0.33	0.14	0.08	0.52
G-H,	0.19	0.11	0.03	0.41	0.31	0.16	0.09	0.50
LFC-L	0.48	0.36	0.05	1.31	0.54	0.34	0.11	1.03
LFC-M	0.73	0.45	0.25	1.66	0.85	0.35	0.17	1.33
LFC-H	1.36	0.97	0.19	3.47	0.69	0.41	0.20	1.33

Related to first-order Ambisonics decoding, the results presented here derive from a B-Format basic decoder. Some analyses were made from measurements using other first-order Ambisonics decoders as in-phase and max-rE/in-phase. The analyses of the acoustics indices showed that using in-phase decoding the LFC values are much lower compared to the reference sound field, resulting in a much greater JND difference in the whole frequency range.

6. CONCLUSIONS

An objective comparison between a SIRR-based and a first-order Ambisonic basic decoder was presented. SIRR-based decoding explained here, uses CQT representation for frequency-time analysis. From diffuseness estimation, the signal is divided in non-diffuse and diffuse sound. Non-diffuse sound is rendered via VBAP method and a maximum of three loudspeakers using the direction information of the instantaneous intensity vector. Diffuse sound is reproduced in all speakers by means of a series of modified reciprocal MLS signals.

Compared with the reference sound field, results showed that both decoders, using JND criteria, have comparable values concerning the reverberation profile and the acoustic indices EDT, Ts, C80, G. However, it was observed that SIRR-based decoding offers narrower and more accurate direct sound reproduction than first-order Ambisonics basic decoding. In the same way, SIRR-based decoding presents lower values regarding the IACC (Late) in the high frequencies. Furthermore, concerning LFC values, it was found that SIRR-based reproduction is more comparable to the reference sound field in the high frequency band than first-order Ambisonics. As is well known, both parameters (IACC and LFC) are relevant for room spaciousness.

From the previous results we can expect that SIRR-based decoding improves the sound incidence and the spatial impression reproduction from B-Format room impulse responses, compared to first-order Ambisonics basic decoding. That is, from an objective point of view, SIRR-based reproduction gives a better 'room impression' which is primordial in the context of concert hall auralization. This observation needs confirmation for other first–order Ambisonics decoders. The next step is formal listening tests to assess the improvements.

ACKNOWLEDGEMENTS

Thanks to professor Ning Xiang for suggesting the use of, and providing, the reciprocal MLS signals. This work is part of a doctorate thesis supported by Los Andes University in Bogotá, Colombia.

REFERENCES

1. Kleiner M, Dalenbäck B, Svensson P. Auralization-an overview. Journal of the Audio Engineering Society. 1993;41(11):861–875.

2. Møller H. Fundamentals of binaural technology. Applied acoustics. 1992;36(3):171–218.

3. Daniel J. Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia. Pierre et Marie Curie; 2001.

4. Guastavino C, Katz BFG. Perceptual evaluation of multi-dimensional spatial audio reproduction. The Journal of the Acoustical Society of America. 2004;116(2):1105.

5. Bertet S, Daniel J, Parizet E, Warusfel O. Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources. Acta Acustica united with Acustica. 2013 Jul;99(4):642–657.

6. Merimaa J, Pulkki V. Spatial impulse response rendering I: Analysis and synthesis. Journal of the Audio Engineering Society. 2005;53(12):1115–1127.

7. Berge S, Barrett N. High angular resolution planewave expansion. In: Proceedings of the 2010 Ambisonics Symposium; 2010.

8. Merimaa J. Analysis, synthesis, and perception of spatial sound: binaural localization modeling and multichannel loudspeaker reproduction. Helsinki University of Technology. Espoo; 2006.

9. Figueiredo F. Indices acoustiques et leurs rapports statistiques : vérification objective et subjective pour un ensemble de salles de spectacles. Pierre et Marie Curie; 2011.

10. Xiang N. Personal communication. 2013

11. Griesinger D. The importance of the direct to reverberant ratio in the perception of distance, localization, clarity, and envelopment. In: Audio Engineering Society Convention 126. Audio Engineering Society; 2009.

12. Pulkki V. Virtual sound source positioning using vector base amplitude panning. Journal of the Audio Engineering Society. 1997;45(6):456–466.

13. Schörkhuber C, Klapuri A. Constant-Q transform toolbox for music processing. In: 7th Sound and Music Computing Conference, Barcelona, Spain; 2010. p. 3–64.

14. Schörkhuber C, Klapuri A, Holighaus N, Dörfler M. A Matlab Toolbox for Efficient Perfect Reconstruction Time-Frequency Transforms with Log-Frequency Resolution. In: Audio Engineering Society Conference: 53rd International Conference: Semantic Audio. Audio Engineering Society; 2014.

15. Brown JC, Puckette MS. An efficient algorithm for the calculation of a constant Q transform. The Journal of the Acoustical Society of America. 1992;92(5):2698–2701.

16. Del Galdo G, Taseska M, Thiergart O, Ahonen J, Pulkki V. The diffuse sound field in energetic analysis. The Journal of the Acoustical Society of America. 2012;131:2141.

17. Xiang N, Trivedi U, Oh J, Braasch J, Xie Bs. Adapting spaciousness of artificial, enveloping reverberation in multichannel rendering based on coded sequences. Proceedings of Meetings on Acoustics. 2013 Jun;19(1):015035.

18. Espitia H JP, Dujourdy H, Polack J. Caractérisation expérimentale du microphone SoundField ST250 pour la mesure de la diffusivité du champ sonore. In: Actes du 12e Congrès Français d'Acoustique. Poitiers; 2014. .

19. Harker A, Tremblay PA. The HISSTools Impulse Response Toolbox: Convolution for the Masses. In: Marolt M, Kaltenbrunner M, Ciglar M, editors. Proceedings of the International Computer Music Conference. The International Computer Music Association; 2012. p. 148–155.