Auditory perception stability evaluation comparing binaural and loudspeaker Ambisonic presentations of dynamic virtual concert auralizations

David Thery, Brian F. G. Katz

JASA ARTICLE

Auditory perception stability evaluation comparing binaural and loudspeaker Ambisonic presentations of dynamic virtual concert auralizations

David Thery[1,a] and Brian F. G. Katz[2,b]

[1]*Laboratoire d'Informatique Pour la Mécanique et les Sciences de l'Ingénieur, LIMSI-CNRS, UPR 3251, Université Paris-Saclay, Centre National de la Recherche Scientifique, Orsay 91400, France*

[2]*Sorbonne Université, Centre National de la Recherche Scientifique, Institut Jean Le Rond d'Alembert, Unité Mixte de Recherche 7190, Lutheries–Acoustique–Musique, Paris, France*

ABSTRACT:

Auralizations can be computed in a variety of ways as well as be rendered over different sound reproduction systems. They are used as a design tool in architectural projects and for fundamental studies on spatial perception and cognition, hence requiring reliability and confidence in the obtained results. This study assessed this reliability through auditory perception stability by comparing the perceived differences between two rendering systems for a given set of second-order Ambisonic auralizations: virtual loudspeaker binaural rendering over head-tracked headphones versus 32-loudspeaker rendering. Anechoic extracts of jazz pieces have been recorded and presented in various acoustic conditions over these two systems, evaluated on the following criteria: Readability, distance, listener envelopment (*LEV*), apparent source width (*ASW*), reverberance, and loudness. Results show that consistent significant differences between scene conditions are comparably perceived across the two systems. However, significant effects of the sound reproduction system were observed for *ASW*, *LEV*, and reverberance in some configurations. © *2021 Acoustical Society of America*. https://doi.org/10.1121/10.0002942

(Received 12 May 2020; revised 18 November 2020; accepted 30 November 2020; published online 11 January 2021)

[Editor: Francesco Martellotta] Pages: 246–258

I. INTRODUCTION

A. Context and problematic

Auralizations have reached a certain level of maturity (Postma and Katz, 2016) and are used in a variety of applications, from virtual archaeological reconstruction to spatial cognition studies to acoustic design in architectural projects for design decision-making. While auralizations are more often presented over headphones, which are convenient due to their portability, easy access, and low cost, there are instances and circumstances when it can be useful to have several people listening to the same auralization, letting them experience together the simulation. The use of loudspeaker-based systems is therefore better for this case, although other reasons such as listener comfort or aesthetics of the listening room are valid ones for selecting a speaker-based system rather than headphones rendering (Thery *et al.*, 2019).

Ideally, the sound reproduction system should not impact the perception of auralizations to ensure that reliable design decisions are made through stable auditory perception.

B. Binaural Ambisonic and loudspeaker rendering

Gerzon (1975) introduced the concepts of what we call today first-order Ambisonic (FOA) recording and playback technology. Ambisonic is a format that allows for the representation of a 3D-sound field through the use of spherical harmonics (SH) (Nicol, 2010). This format, agnostic to rendering configuration, allows easy spatial manipulations, such as rotation, directional loudness control, and warping (Alary *et al.*, 2019; Zotter and Frank, 2019). Ambisonics provides a decoupling between encoder and decoder. The encoder is purely linked to the SH, while the decoder is defined by the loudspeaker arrangement. In this way, the number of loudspeakers is independent of the number of encoded virtual sources (Noisternig, 2003).

Ambisonic encodings can be decoded to any loudspeaker layout as well as to headphones (binaural Ambisonic), with rendering over headphones typically employing a virtual speaker array approach. This approach is based on the decoding of the Ambisonic stream to virtual loudspeaker positions, from which the binaural signals are then created through binaural rendering via convolution with the head-related transfer function (HRTF) appropriate to their spatial positions. These individual processed speaker signals are then summed to create the left and right ear headphones signals (Jot *et al.*, 1999; Noisternig, 2003).

An active field of research for improving binaural renderings concerns the *HRTF*, leading to the recent standardization of these filters with the *SOFA* format (Majdak and Noisternig, 2015). Research in binaural audio aims at improving/correcting localization, externalization, coloration effects, and spatial impression when performing

[a]Electronic mail: david.theryfr@gmail.com, ORCID: 0000-0001-9431-0605.

[b]ORCID: 0000-0001-5118-0943.

https://doi.org/10.1121/10.0002942

renderings over headphones. The individual nature of the HRTF is clearly an active field of research, from 3D modeling/scans of ear morphology to machine-learning based approaches to perform HRTF individualization. A related field to note is the improvement of the calculation of the headphone transfer function (HpTF) (Engel *et al.*, 2019; Paquier and Koehl, 2015).

Similarly, Ambisonic rendering over loudspeakers has gained increased attention due to general hardware/software support for multichannel audio, starting its democratization in the sound engineering community, and recently concisely reviewed in Zotter and Frank (2019). Research studies have focused on limiting coloration effects, extending the *sweet spot* area to off-center positions (Stitt *et al.*, 2017), or improving the localization accuracy and spatial impression through the development of various decoding methods (Zotter *et al.*, 2013). The effect of the setup room's acoustics also has major influence in Ambisonic rendering, and hence needs to be controlled (Lokki, 2011). Frank (2014) discussed the main factors influencing *localization*, *source width*, *coloration*, and *loudness*, including the *number of speakers*, *Ambisonic order*, *array radius and compensation filters*, *decoding strategy*, *reproduction room*, and *reverberation time*. He concluded that while a large number of loudspeakers achieve good localization for the correspondingly suitable order, order truncation (or correspondingly too many speakers) can result in coloration issues. Using an insufficient number of speakers (i.e., too high an order for the corresponding loudspeaker array) can also cause imbalanced timbre, source width, and loudness. The regularity of the loudspeaker layout is also of crucial importance for sound field accuracy, no matter the employed decoding method. Severe coloration can be induced by delay compensation filters as well, which does not improve localization at the center (Stitt *et al.*, 2017).

C. Related work

Very few studies have compared binaural to loudspeaker rendering of Ambisonic content. There have been studies concerning binaural and loudspeaker playback, outside of Ambisonic usage, and other studies concerned with Ambisonic rendering variances.

Fischetti *et al.* (1993) compared dummy-head recorded samples rendered over headphones and a two loudspeaker stereo system in an acoustically damped studio (evaluations temporally spaced by ten days). The dummy-head recordings were originally conducted in an acoustically modular space (the "Espace de projection" at IRCAM) in a variety of acoustic configurations (position in the room, ratio of absorptive to diffusing panels, and ceiling height). The musical stimulus was a 15 s piece of Schubert's 14th string quartet. Ten sound engineering students evaluated samples according to six attributes: *apparent room size, depth perception (or relative distances), lateral localisation*, *spatial impression*, and *reverberance*. Their results showed that the effect of the reproduction system was particularly relevant for distant recorded positions: subjective reverberation time (RT) was longer and depth perception poorer over loudspeakers. It was also reported, for loudspeaker presentation across positions: (1) values of spatial impression were lower, (2) inter-individual variance was higher, and (3) larger values of apparent room size were more correlated with high ceiling with absorptive configurations. Results were interpreted as a less accurate representation of diffuse field spatial characteristics through loudspeakers, mainly due to the source positions.

Guastavino and Katz (2004) compared different loudspeaker configurations using Ambisonic recordings, ranging from urban soundscapes to musical excerpts. Twenty-six expert listeners were involved in the evaluation of FOA recorded soundscapes rendered over one-dimensional (1D) (2.1), 2D (6.1), and 3D (12.1) loudspeaker systems (the 1D comprised solely a stereo pair, the 2D comprised six speakers on a circle around the head of the listener at ear height, and the 3D was a hexagonal structure based on the 2D array), in a small damped room (RT less than 0.05 s above 200 Hz, gradually increasing to 0.2 s at 40 Hz). The evaluated parameters were *readability, presence, distance, localization, coloration*, and *stability*, while a preliminary experiment also evaluating *naturalness* and *immersion*. Results exhibited significant differences between 1D, 2D, and 3D arrays, and in particular:

- 2D was evaluated as providing a higher degree of readability, more immersive, and closer than 3D array; 1D was judged even less immersive and farther away.
- A correlation between choice of the most "natural" method and the specific soundscape (1D, 2D, and 3D reproduction methods were, respectively, more adapted to frontal musical scenes, outdoor environments, and indoor environments). This correlation was also present for coloration, localization, and distance (linked to "natural"). An evident correlation between choice of the most natural method and the specific soundscape.
- A strong correlation between readability and localization for all three reproduction methods ("sources can be easily located in a spatially well-defined environment"), as well as between presence and distance ("an immersive scene sounds close").
- Concerning localization and coloration, the 3D reproduction was perceived as indistinct and muffled in comparison to the 1D and 2D reproductions, which were described as clearer and more precise.

To summarize, there was a noticeable difference between 1D, 2D, and 3D in terms of perceived distance, presence, and stability. The judgments for the 3D representation fall between the 1D and 2D method values for all parameters but coloration. This agrees with Frank (2014), who reported an increase in coloration with increasing number of loudspeakers. It also echoes Marentakis *et al.* (2014), where the conditions with fewer loudspeakers were mostly preferred.

Guastavino *et al.* (2007) compared three recording techniques with associated sound reproduction systems, namely:

(1) stereo (ORTF) recordings, played back with two loudspeakers located in front of the listener at +/− 30° azimuth; (2) dummy-head recordings, played back using the same system with transaural processing; and (3) B-format (FOA) recordings, played back with six loudspeakers regularly spaced around the listener, including the two speakers used in other conditions. Eleven expert listeners evaluated subjective parameters including *envelopment*, *immersion*, *representation*, *readability*, *realism*, and *overall quality*. The recorded scenes included outdoor traffic noise and three indoor recordings (car interior, people talking with background music, and an excerpt of electric guitar). A significant difference was observed between transaural and both Ambisonics and stereo for the concert excerpt (only one in four sound scenes), while being consistent for the remaining scenes. When considering all sound scenes, significant effects of reproduction techniques were observed for *envelopment*, *immersion*, *readability*, *realism*, and *overall quality*. In particular, Ambisonics was rated as significantly more enveloping, more immersive, as well as significantly less readable than both transaural and stereo. Regarding overall quality, stereo and Ambisonics were rated significantly higher than transaural. An additional experiment in this paper showed in particular that transaural was judged precise and as providing easy localization and good readability, while lacking realism and immersion/envelopment; FOA provided strong immersion and envelopment while lacking in localization and envelopment; stereo rendering provided very precise localization while lacking envelopment.

Koehl *et al.* (2011) compared the subjective ratings of expert listeners (12 sound engineering students), comparing headphones (without binaural processing) versus loudspeaker setups, including frontal mono, stereo pair (±30°), and ITU 5.1 setup, using various recording methods (from mono to stereo to multi-channel). Their study aimed at evaluating whether differences between sound sequences were equally perceived when played back over headphones as over loudspeaker systems, through the rating of *similarity* of stimuli pairs (from "identical" to "extremely different") successively rendered over different sound systems. They concluded that whatever the audio content (mono, stereo, multi-channel), the differences between the two recording systems were equally perceived with headphones as with loudspeaker setups, providing good consistency across systems. Also, the headphones condition led to better consistency between listeners, meaning larger variations were observed with the loudspeaker setups. It should be noted, however, that the attribute *similarity* may lack discriminability, as it encompasses several dimensions, such that listeners could rate it using different strategies (e.g., similar in timbre vs spatial impression).

These most relevant studies have shown that sound reproduction system can have a significant impact on subjective evaluation of variously recorded scenes, whether concerning timbral (e.g., through coloration) or spatial attributes (envelopment, immersion). In addition, it was shown to be stimuli-content-dependent.

Simulated auralizations have reached a high degree of authenticity when compared to measured ones. With this increased ecological validity, the present study aims at comparing binaural Ambisonic to Ambisonic over loudspeakers, rendering simulated auralizations, with a focus on spatial room acoustic attributes. The proposed hypothesis is that significant differences between acoustic configurations are consistently perceived across Ambisonics rendered binaurally over headphones and the same Ambisonics rendered over loudspeakers.

Section II presents the creation of the auralizations, including anechoic recordings and geometrical acoustic (GA) models descriptions. The experimental design is described in Sec. III, followed by the results in Sec. IV, discussed in Sec. V, leading to the conclusion in Sec. VI.

II. AURALIZATIONS

This section describes the creation of the auralizations used as stimuli in the present experiment, including anechoic recordings, room model creation and calibration, and inclusion of dynamic source directivity for movable instruments.

A. Anechoic sources

Anechoic jazz extracts have been selected from a recently published anechoic audio and 3D-video content database (Thery and Katz, 2019). Two jazz trio extracts of different styles/periods were employed; the criteria were to have different tempo, as well as different orchestrations, while having at least one moving instrument to highlight effects of dynamic source directivity, a feature available in the database. The extracts used were:

- Django Reinhardt: Minor Swing, interpreted by Double-Bass, Guitar, Violin
- Sydney Bechet: Si tu vois ma mère, interpreted by Double-Bass, Guitar, Saxophone alto

The detailed procedure of these anechoic recordings is reported in Thery and Katz (2019). To summarize, recordings included two sessions, after some rehearsals all together: the first one in which all musicians were recorded playing together, to serve as a reference for the second one in which each musician was recorded individually, while listening to the reference recording from which their own instrument was removed. Close-microphones (DPA-4060) mounted on the instruments were used for both moving instruments violin and saxophones, while fixed omnidirectional microphones (DPA-4006) were used for the guitar, cello, and double-bass, placed approximately 1 m away from the source. In order to provide a natural acoustic for the musicians, the live audio feed of each instrument was processed in real time, adding a constant reverberation of 1 s, rendered identically to all musicians over open headphones (Sennheiser HD650), to simulate a small recording studio environment using the SPAT (Carpentier, 2015). The virtual monitor in SPAT was placed in front of the listener,

https://doi.org/10.1121/10.0002942

with a wide aperture of 90°, a room size of 600 m^3, and reverberation with a flat spectrum. Playback level was adjusted individually for each musician to optimize playing comfort.

In parallel to the audio recordings, depth-video recordings were performed using a multiple Kinect v2 system based on the LiveScan3D library (Kowalski *et al.*, 2015). Three Kinect sensors were used to capture musicians RGB-D videos, subsequently enabling to reconstruct point-clouds in a VR scene, as detailed in Poirier-Quinot *et al.* (2016). These video recordings enabled the extraction of the orientation of the moving instruments with a tracking object that followed the instrument movement, to be used as input for the incorporation of dynamic source directivity.

B. Rooms

In order to keep the focus on realistic listening conditions, rather than simplified laboratory impulse sequences, two small similar sized music halls suitable for the music stimuli were selected, representing two different geometries: a shoe-box hall, and a semi fan-shaped hall [see Barron (2009) for a detailed review of the various shapes of halls in architectural acoustics, and their acoustic objective and perceptual consequences]. The first was the Morgan Museum Library (MML) auditorium in New York, by architect *Renzo Piano* (Renzo Piano Building Workshop), and acoustics by *Kahle Acoustics*, for which a detailed study of the acoustic design is available (Katz and Kahle, 2008). This hall has 300 seats and a volume of 2000 m^3, with the majority of the walls covered with reflecting panels as a well as hanging ceiling shaped panels. The second was the Amphitheater of the Cité de la Musique (CM) in Paris, by architect *Christian De Portzamparc* and the acoustics handled by *Commins BBM, ACV, and Xu Acoustique*. This hall has 230 seats and a volume of 1370 m^3.

1. GA model creation and calibration

GA room model creation and calibration was performed using the GA software CATT-ACOUSTIC (v9.1, TUCT v2.0). Initial simulations were performed with algorithm 1, while final RIR simulations were computed using algorithm 2. Algorithms 1 and 2 differ in the way diffuse rays are reflected. Two methods are used: deterministic split ray and random scattering. Algorithm 1 uses random scattering with optional ray splitting for up to second-order reflections. With random scattering, each incident ray generates a single reflected ray, either specular or scattered, with a probability depending on the scattering coefficient, thus requiring a large number of rays, as no additional rays are created. Too few rays may result in insufficient reflection density in the late part of the decay to be suitable for auralization. Besides, the late impulse response may vary significantly between successive calculation runs, due to the randomness of the angle of reflection of diffuse reflections, potentially leading to large variations in T30 between successive runs. In contrast, algorithm 2 creates for each incident ray a specular reflection and many new rays representing diffuse reflections (with low energy), resulting in an increase in reflection density as the sound decays. Algorithm 2 is much more computationally expensive, but generally yields much more accurate results as well, particularly for complex rooms (Dalenbäck, 2018).

The GA model of the MML was created previously (Katz and Kahle, 2008) and is reused here. The GA model of the CM was created based on 2D plans provided by the *Cité de la Musique*, photographs, and on-site measurements. They are depicted in Fig. 1. One important aspect to mention in the creation of GA models is its level of detail (LOD). This parameter impacts both the computational time of the simulations and the accuracy of the predictions (Dalenbäck, 2018; Savioja and Svensson, 2015). It has been reported that differences of structural details below the size of 70 cm become more and more indistinguishable, and the simplification of the GA model of CM was therefore based on this value (Pelzer *et al.*, 2010; Savioja and Svensson, 2015).

In order to focus attention on the differences in spatial attributes between rendering systems, it was decided to attempt to limit the variations of reverberation time between the two room. The GA model of CM was adjusted to match the T30 of MML. Similarly to the GA calibration procedure described in Postma and Katz (2016), this step was performed by adjusting the absorption of defined materials until the difference in mean (T30) between rooms was below just noticeable difference (JND), which is 5% for T30 (ISO, 2009), for ten consecutive runs, taking into account the run-to-run variation of GA algorithms with statistical scattering implementations.

Table I presents the final GA models acoustic parameters T30, EDT, and C80, for two receivers (front and rear seating positions), for the central source described below, in both rooms, after calibration, and for octave bands from 125 to 4000 Hz). Concerning perceptual relevance for some parameters, it is more reasonable to consider mid-frequency average values (average of octave-band results for 500 to 2000 Hz. These values are shown in Table II, which presents averaged values for early and late interaural cross correlations ($IACC_3$) and direct-to-reverberant ratios (DRR_3), that are relevant to support results analysis, for spatial attributes (see Secs. IV B and IV C). In particular, $IACC_{E3}$ and $IACC_{L3}$ are, respectively, strongly correlated with the perception of ASW and LEV. It is noted that the JND for IACC is 0.075 (ISO, 2009). Similarly, the DRR is one of the two main cues for the perception of distance (with sound level), and is also related to the perception of reverberance. The generally accepted JND for this parameter is $JND_{DRR} \approx 5$ dB (Zahorik, 2002), while being variable with absolute level.

Figure 2 depicts spatio-temporal visualizations of RIRs that provide more details on the distribution of reflections (both in terms of time and orientation), performed using the spatial decomposition method (Tervo *et al.*, 2013).

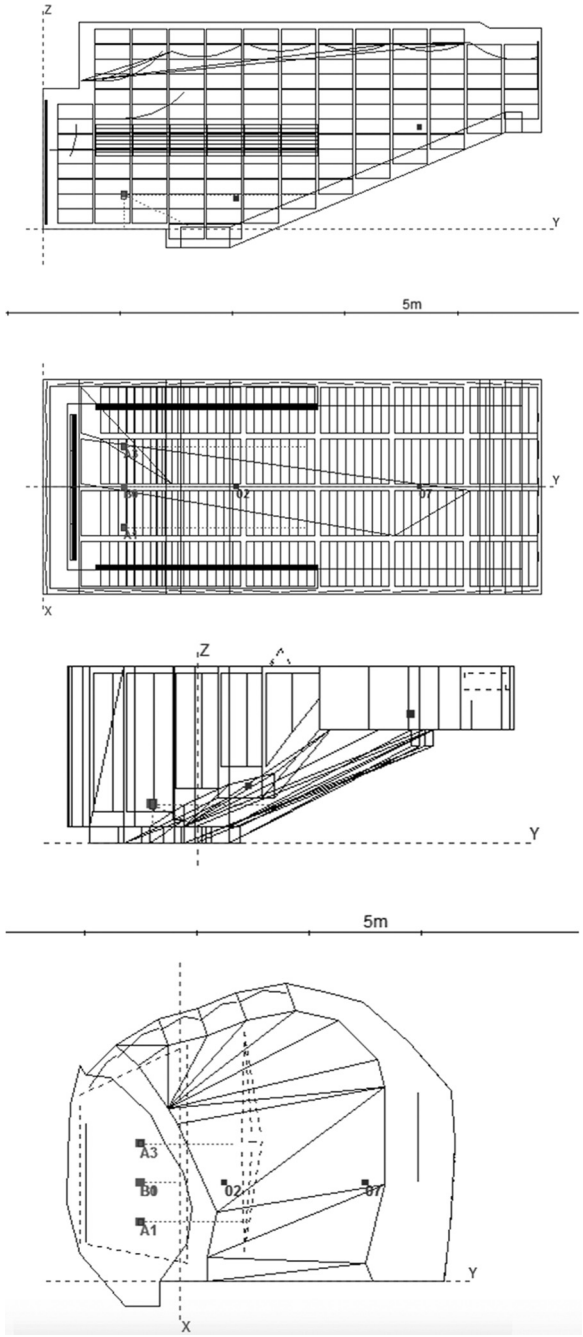

FIG. 1. GA models of Morgan Museum Library (top), and Cité de la Musique (bottom), with sources and receivers position.

2. RIR simulations and convolutions

The sources representing the instruments were placed as follows: the moving (rotation only) instrument (saxophone or violin) placed at the center of the stage, the double bass on the right when looking at the stage, the guitar on the left, both equally spaced by 2 m in both rooms. Two receiver positions (front and back) were defined to have two significantly different DRR values for providing different acoustic conditions. The receivers distance from the central source were chosen to have equal ratios across rooms between each S-R and total length of the given room (with S being the central source, i.e., the moving instrument), as represented in Fig. 1.

Once the GA models have been made and calibrated, detailed RIR simulations were performed for the auralizations. For the static instruments, directivity patterns were applied, respectively, for the double-bass and the guitar, provided in CATT. Simulations were run (algorithm 2) and exported as second-order Ambisonic RIRs, for each S-R pairs in both rooms. The second-order Ambisonic RIRs were then convolved with each instruments' anechoic extract.

To take into account the dynamic directivity (variable orientation) of the center source, a composite source was defined, following the method presented in Postma *et al.* (2016). This study has shown that simulated auralizations including dynamic voice directivity are perceived as more plausible, more enveloping, and wider in terms of *ASW*, than with static (frontal) directivity. This approach is based on the decomposition of an omni-directional RIR into 12 equally distributed beam patterns. The beams are designed to have minimal overlap while keeping an equal gain sum in order to approximate an omni-directional pattern. Dynamic orientation of the instruments were extracted from the *Kinect* video recordings, allowing the incorporation of dynamic directivity in the auralizations of these moving instruments, as detailed in Postma *et al.* (2016). Another source of dynamic directivity are variations between notes, as shown in Shabtai *et al.* (2017). Although it could be interesting to study this additional factor, it was not implemented in the present simulations.

III. EXPERIMENTAL DESIGN

The experiment consisted of the presentation of eight different stimuli, providing all combinations of the two musical pieces, two rooms, at two listening positions, each repeated three times.

A. Stimuli selection

One of the main constraints when conducting listening tests is the duration of the experiment, which can lead to subject fatigue and overall results bias (Zacharov, 2019). To allow for testing a wider range of acoustic conditions, the stimuli presented here were short musical passages. A maximum overall test duration of 1 h was targeted. Considering the eight conditions (two music extracts, two positions, two rooms), three repetitions, and two sound reproduction systems, extracts of ≈20 s were deemed appropriate. The two extracts from *Minor Swing*, labelled *Django*, and *Si tu vois ma mere*, labelled *Bechet*, were, respectively, 20 and 18 s in duration. These two extracts provided two musical phrases at two different tempos [respectively, 200 and 80 beats per minute (BPM)], with slightly different instrument arrangements: violin vs saxophone, with double-bass and guitar present in both extracts.

https://doi.org/10.1121/10.0002942

TABLE I. T30, EDT, C80, IACC, and DRR (direct-to-reverberant ratio) values from the RIRs analysis, both for MML and CM. Values for centered source, front/back receiver positions (respectively, receiver 02 and 07 in Fig. 1).

Room	Octave band	125	250	500	1000	2000	4000
MML	T30	1.05/1.20	1.10/1.14	1.17/1.16	1.16/1.23	1.13/1.02	1.08/0.91
	EDT	1.20/ 0.98	1.13/1.10	1.34/1.05	1.02/1.09	0.99/1.05	0.85/1.03
	C80	−0.67/4.01	1.94/2.55	0.93/2.52	0.92/1.14	3.00/3.34	2.69/1.14
	$IACC_{early}$	0.97/0.95	0.95/0.91	0.86/0.73	0.58/0.29	0.25/0.41	0.23/0.42
	$IACC_{late}$	0.94/0.96	0.90/0.83	0.67/0.72	0.47/0.48	0.65/0.52	0.58/0.44
	DRR	−4.9/−2.9	−3.9/−4.4	−0.3/−5.1	−1.5/−2.0	−2.1/−4.0	−4.8/−3.0
CM	T30	1.08/1.16	1.10/1.13	1.17/1.14	1.20/1.16	1.17/1.05	1.04/1.08
	EDT	0.57/1.39	0.70/1.24	0.86/1.31	1.19/1.31	0.96/1.26	0.93/1.12
	C80	7.02/−0.16	5.35/−2.76	4.12/−0.65	1.60/−1.22	2.86/−1.09	4.21/−1.01
	$IACC_{early}$	0.98/0.96	0.85/0.87	0.68/0.52	0.35/0.16	0.34/0.46	0.40/0.32
	$IACC_{late}$	0.97/0.95	0.83/0.87	0.60/0.58	0.30/0.31	0.50/0.57	0.33/0.44
	DRR	−1.4/−3.1	−3.5/−4.2	−6.3/−7.3	−3.7/−5.1	−6.4/−6.2	−4.4/−6.5

B. Auralization rendering

The second-order Ambisonic files obtained after convolution were rendered both binaurally and over a 32-speaker loudspeaker array, using SPAT, set with *HOA3D, Energy-preserving, MaxRE* parameters.

The binaural rendering was performed using the virtual speaker array approach, with *KEMAR HRTF* and 18 virtual speakers, preferred for its flexibility over a more recent method proposed in Zaunschirm *et al.* (2018). 18 speakers were used for the second-order Ambisonic rendering, with the main concern being on the optimal distribution of virtual speakers (as uniform as possible over a sphere's surface), to avoid ill conditioning or singularities in the decoder matrix (Noisternig, 2003). Informal perceptual evaluation of the rendering was judged consistent between systems.

While Frank (2014), Solvang (2008), and Zotter and Frank (2019) have reported that using a too high number of loudspeakers with regards to the number of virtual loudspeakers could be detrimental (mainly producing coloration, and even comb-filtering effects for 2D systems), any potential coloration arising from this choice is addressed through a cross-system equalization step (see Sec. III C). It should be noted that the impairments reported in Solvang (2008) have not yet been validated subjectively. In addition, exact listener head position, head movements, headphones placement, and application of HpTF were not considered as potential variability factors in this study, thereby focusing on the overall difference between systems for various acoustic conditions.

TABLE II. Averaged (avg.) values of DRR and IACC over 500 Hz to 2000 Hz octave bands (DRR_3 and early and late $IACC_3$), both for MML and CM, and front / back positions (pos.).

Parameter	Room	Front	Back	Avg. pos.
$IACC_{E3}$	MML	0.56	0.48	0.52
	CM	0.46	0.41	0.45
	Avg. rooms	0.52	0.45	0.48
$IACC_{L3}$	MML	0.60	0.57	0.59
	CM	0.46	0.49	0.47
	Avg. rooms	0.54	0.51	0.52
DRR_3	MML	−1.3	−3.7	−2.5
	CM	−5.5	−6.2	−5.9
	Avg. rooms	−3.4	−4.5	−4.2

C. Hardware and equalization

Head-tracked binaural rendering was presented over open reference headphones (Sennheiser, HD650), no HpTFs were applied. Head orientation was provided by a ten camera *OptiTrack* tracking system, communicated to Max via OSC.

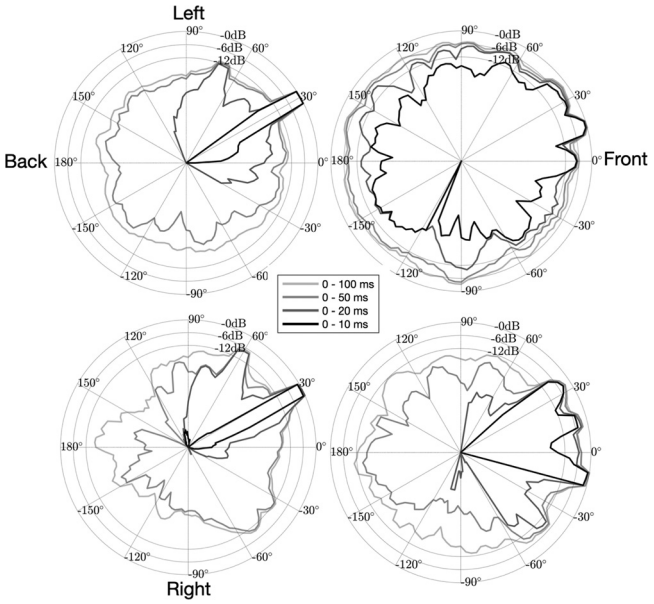

FIG. 2. Spatio-temporal representations (horizontal plane) of the simulated RIRs for CM and MML at front and back positions, for source A1. Top left/right: CM front/back. Bottom left/right: MML front/back. Temporal steps include 10, 20, 50, and 100 ms. The indication *Front* designates the stage of the hall, and *Back* the rear. Performed with the SDM toolbox (Tervo *et al.*, 2013) on the B-format simulated RIRs, band-pass filtered from 100 to 5000 Hz, with an angular averaging smoothing filter of 3°.

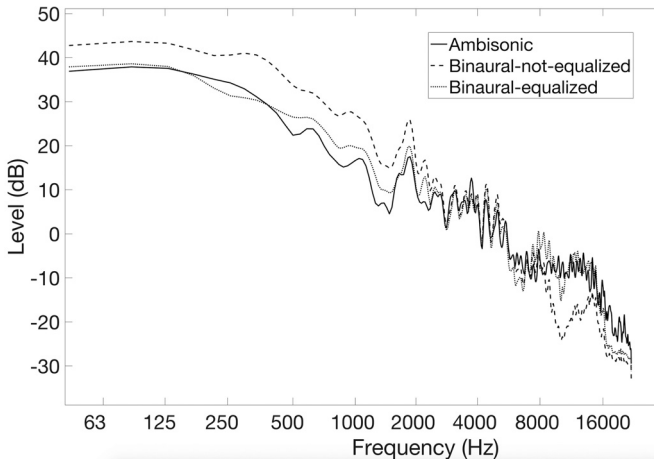

FIG. 3. Average frequency spectra over the entire recording of one stimulus (Bechet), including Ambisonic, and binaural pre and post-equalization.

The Ambisonic loudspeaker system was composed of 32 time-aligned loudspeakers (Model ELAC 301) positioned at three height levels: 8 speakers at floor height, 12 speakers at head height (1.4 m), and 12 speakers around the ceiling. The room, with dimensions 4 m L, 3.6 m W, and 2.6 m H, was acoustically treated, providing a background noise of 20 dB(A), and a reverberation time of 0.1–0.3 s over the 125 to 8000 Hz frequency range.

To reduce the effect of coloration due to small discrepancies between the frequency responses of the two systems (Ambisonic binaural decoding, room acoustics influence), a cross-system equalization step was performed, illustrated in Fig. 3. Binaural recordings on a dummy-head *Neumann KU80* of one room configuration (room and position) for the two musical extracts were made for both presentation systems. The average frequency spectrum of both recordings were computed in octave bands, with the mean value taken between the two channels of the dummy-head and the two musical extracts (see Table III). The resulting differences by octave band between the two rendering systems were then used to generate a second-order bi-quad compensation filter applied to the binaural stream (125 to 8000 Hz octave-band). No normalization between musical extracts were applied, as it would have introduced a timbral modification of the extract. The playback level was calibrated to be on average 75 dB(A) with a binaural recording of one of the stimuli.

D. Attributes selection and evaluation

Following previous research on the comparison of auralizations (Postma and Katz, 2016), six attributes have been selected for this listening test. However, *plausibility* has been replaced by *readability* in the present experiment, as several instruments, at slightly different positions, were playing simultaneously, and being able to hear distinctively (musical clarity) the different voices was deemed more appropriate.

- *Readability*: the ability to focus on a given component of the sound, or to discriminate the different sources, from *very blurred* to *very clear*.
- *Distance*: the perceived acoustic distance from the listening position to the sound scene, from *very close* to *very far*.
- *ASW*: the extent of the sound source on an horizontal plane, from *very narrow* to *very wide*.
- *LEV*: the sensation of being surrounded by the sound, from *not enveloping* to *very enveloping*.
- *Reverberance (Rev)*: the perceived amount of reverberation, from *not reverberant* to *very reverberant*.
- *Loudness*: the perceived loudness, from *very weak* to *very loud*

The evaluation interface was presented on an Apple iPad Pro, mirroring a Max patch, using the Mira toolbox [supporting Open Sound Control (OSC) communication]. Each of the six attributes were presented simultaneously with a corresponding discrete seven-point rating scale. More details about the interface can be found in Thery (2020).

TABLE III. Absolute level difference between dummy-head Ambisonic and binaural recordings, pre and post-equalization of the binaural rendering, for octave bands from 125 to 4000 Hz.

Octave band	125	250	500	1000	2000	4000	8000
Diff. pre-equalization (dB)	4.9	7.4	9.8	10.7	7.3	1.0	2.4
Diff. post-equalization (dB)	−0.8	−2.1	2.1	3.8	1.9	0.1	1.3

E. Procedure

Prior to the tests, participants were asked to fill a consent form to confirm their participation. To ensure a good understanding for all participants, written instructions were provided containing a description of the test procedure and user interface, as well as the definitions of the assessed acoustic attributes. Oral instructions and potential answers to questions followed to ensure all participants had well understood the task and the definition of all attributes. Next, participants went through a familiarization phase, where they could become acquainted with the test interface. During this phase, all eight test stimuli were successively presented over the two rendering systems. Having heard all conditions, participants were encouraged to use the full range of the rating scale, as they had heard all conditions. This familiarization approach is recommended by the *ITU BS.1534–3* (ITU-R, 2015) and is similar to Bech *et al.* (2005). This phase enabled the experimenter to verify if the definitions of all attributes were well understood by each participant. The results of the familiarization phase were not included in the test result analysis.

The experiment was carried out in two sessions, representing the two rendering systems, with a short pause in-between, enabling us to switch the sound reproduction systems and put or remove headphones. Presentation order was randomized, with half of the subjects starting with headphones and half with loudspeaker presentation. Each session comprised 24 stimuli (2 positions × 2 music × 2 rooms × 3 repetitions), randomly ordered. Participants were

https://doi.org/10.1121/10.0002942

able to listen to each stimulus as many times as they needed, although they were recommended to limit to 3 to 4 replays, in order to limit test fatigue and overall duration of the test.

F. Participants

A total of 15 participants took part in the listening test [mean age: 33 years old, standard deviation (σ): 7.5, 1.8 M/F ratio]. All participants had a hearing threshold of less than 20 dB hearing level (HL) for either ear across 125 to 10 000 Hz. A wide range of listening expertise was represented, from naive subjects to expert listeners, all coming from the same laboratory, voluntarily. Participants were not compensated for their participation.

IV. RESULTS

As the experimental design and physical conditions did not allow direct A/B comparison between loudspeaker and binaural rendering, result analysis focused on variation trends between conditions for the two reproduction methods. From one stimulus to another, analysis assessed whether differences across stimuli are similarly perceived between conditions: Ambisonic loudspeaker and binaural decoded headphone. In other words, it is assessed how consistent the perceptual evaluation of the auralizations was across these two systems.

A. Statistical analysis

A first step was the normalization of the responses in order to create comparable ratings between participants, avoiding any potential bias between systems. This normalization is based on the standard score equation (y being the normalized response, x the actual response, μ the mean of the subject's response for the given attribute over all trials, and σ the standard deviation for the same acoustic attribute)

$$y = \frac{x - \mu}{\sigma}. \tag{1}$$

The repeatability of the normalized responses were computed from the absolute difference between the normalized responses across repeated trial conditions, for each acoustic configuration, subsequently averaged. The mean differences between repetitions for each attribute across participants were: *plausibility* = 0.55, *distance* = 0.59, *loudness* = 0.77, *ASW* = 0.45, *LEV* = 0.8, and *reverberance* = 0.41, noting the highest variability across repetitions for LEV, while overall giving confidence in the obtained ratings.

In addition, effect sizes were computed to assess the strength of the differences between conditions, and Cohen's d, d_{Cohen}, are given for each p-value (Cohen, 1990; Cumming, 2014). Cohen (1990) proposed rules of thumb for interpreting effect sizes, suggesting that a value of $|0.2|$ represents a "small" effect size, $|0.5|$ represents a "medium" effect size and $|0.8|$ represents a "large" effect size. A threshold of 0.3 was chosen above which d_{Cohen} differences were large enough to be noted in Table IV (Cohen, 1988, p. 185), based on real-world applications. The computation of d_{Cohen} for paired samples was performed according to Eq. (2),

$$d_{\text{Cohen}} = \frac{\mu - \nu}{\sigma_{\text{Pooled}}}, \tag{2}$$

where μ and ν are, respectively, the means of the first and second compared distributions, and

$$\sigma_{\text{Pooled}} = \frac{\sqrt{\sigma 1^2 + \sigma 2^2}}{2}. \tag{3}$$

In the following, variation trends between loudspeaker and headphone renderings are analyzed by independent variable, namely, position, room, and musical extract. Of interest is whether or not significant differences between conditions are consistently observed for the same test conditions between the two rendering systems.

TABLE IV. *p-values* for Wilcox signed rank tests, related to Figs. 4, 5, and 6. ε represents p-values that are smaller than 0.01. d_{Cohen} effect sizes are reported below each p-value.

Attribute	Position	Room	Musical extract
	$p_{\text{Amb}}/p_{\text{Bin}}$ Front vs back	$p_{\text{Amb}}/p_{\text{Bin}}$ MML vs CM	$p_{\text{Amb}}/p_{\text{Bin}}$ Django vs Bechet
Readability	ε/ε	0.07[a]/0.70[a]	ε/ε
$ES - d_{\text{Cohen}}$	0.52/0.63	−0.22/0.04	−0.75[b]/−0.43[b]
Distance	ε/ε	0.83[a]/0.12[a]	ε/ε
$ES - d_{\text{Cohen}}$	0.09/ε	ε/−0.18	0.76/0.63
Loudness	ε/ε	ε/ε	ε/ε
$ES - d_{\text{Cohen}}$	0.17/0.15	0.36[b]/ 0.83[b]	−1.0/−1.22
ASW	0.94[a]/0.03	7e − 6/ε	ε/ε
$ES - d_{\text{Cohen}}$	0.12/0.11	0.53/0.65	−0.41/−0.45
LEV	5.1e − 5/0.10[a]	ε/ε	ε/ε
$ES - d_{\text{Cohen}}$	0.15/0.13	0.40[b]/0.71[b]	−0.77[b]/ε
Rev.	0.81[a]/0.53[a]	ε/ε	ε/ε
$ES - d_{\text{Cohen}}$	0.17/0.02	0.60/0.76	−0.48[b]/−0.13[b]

[a]Non-significant differences between conditions ($p > 0.05$).
[b]Differences between trends above 0.3 threshold across rendering systems.

B. By position

Results are analyzed by rendering system and position, comparing front and back positions (see Fig. 4). For the loudspeaker condition, all attributes except *ASW* and *reverberance* were rated significantly different (front position being more readable, closer, more enveloping, and louder). For the binaural headphone condition, all attributes except *LEV* and *reverberance* were rated significantly different (front position more readable, closer, louder, and narrower). Consequently, it can be seen that for *ASW*, the two systems exhibit similar trends, although the difference in responses between front and back positions under loudspeaker rendering were not statistically significant. The same is observed for *LEV*, with the distinction between front and back positions not reaching significance levels for the binaural headphone condition ($p = 0.07$, see Table IV). It should be noted that LEV ratings exhibited a large variance (and the related largest mean repeatability), especially in the loudspeaker

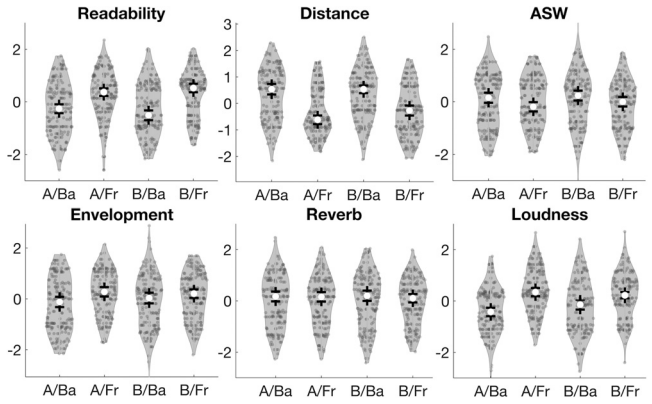

FIG. 4. Distribution of normalized results comparing (A)mbisonic loudspeaker and (B)inaural headphone, for (Fr)ont and (Ba)ck positions. Results are aggregated over all subjects, rooms, and extracts; (○) median, (+) 95% confidence interval limits.

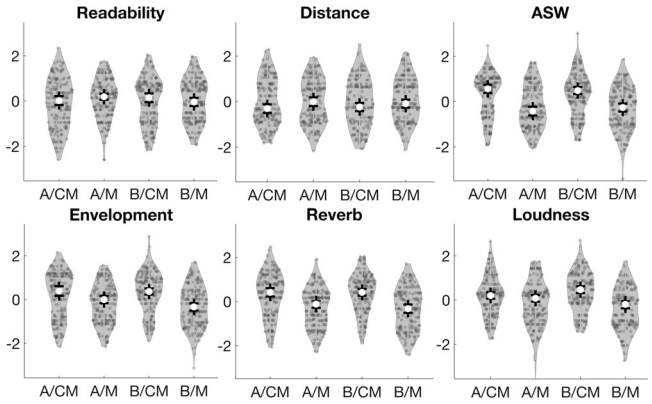

FIG. 5. Distribution of normalized results comparing (A)mbisonic and (B)inaural, for rooms (M)ML and (CM). Results are aggregated over all subjects, positions, and extracts. (See Fig. 4 for plot details.)

condition, which could have influenced this difference between systems.

Effect sizes were similar for all attributes, reinforcing the consistency of judgments across systems. The mean difference across all attributes was $d_{\text{Cohen}} = 0.05$. The largest difference between loudspeaker and headphone being 0.15 for *reverberance*, below the defined threshold of 0.3 (see Table IV).

Regarding the objective RIR parameters, averaged IACC_3 and DRR_3 values presented in Table II do not appear to support the observed results, given the large variance on the evaluation of *ASW* and *LEV*, in addition to the low differences between these averaged values [close to the respective $JND_{\text{IACC}} = 0.075$ (ISO, 2009) and $JND_{\text{DRR}} \approx 5\,\text{dB}$ (Zahorik, 2002)]. However, it can be seen in Fig. 2 that the back positions received more early lateral reflections (especially true for CM), probably explaining the higher ratings of ASW at the back position. These plots also highlight the acoustic energy contribution of the rear of the wall, clearly visible in CM, but not in MML. This could be explained by the smaller size of CM, and the back position being closer to the reflecting surface in this hall.

In summary, responses for *readability, distance, reverberance, and loudness* followed the same trends between rendering systems. In contrast, an effect of rendering system was observed on the evaluation of *ASW* and *LEV*.

C. By room

Results by room condition are presented in Fig. 5. All attributes followed the same trends with both sound reproduction systems: no significant differences appeared for *readability* and *distance* between rooms, while CM was judged significantly wider, more enveloping, more reverberant, and louder than MML.

d_{Cohen} indicated, however, larger differences between rooms for the binaural headphone system regarding *loudness* and *LEV* attributes. Even if the threshold of 0.3 is not reached, CM was judged slightly more *readable* than MML with the loudspeaker system (d_{Cohen} difference = 0.26).

These results (in particular *reverberance*, but also *ASW* and *LEV* which are positively correlated with *Rev*) are supported by the direct-to-reverberant ratio values presented in Table I, which show lower DRR values for CM than MML in the mid-range (500, 1000, and 2000 Hz octave bands). Regarding other objective RIR parameters, IACC_3 early and late values are both lower in CM than in MML (with differences above of the respective IACC_3 and DRR_3 JNDs), which is in agreement with the significantly higher ratings of ASW (for $\text{IACC}_{\text{early}}$) and LEV (for $\text{IACC}_{\text{late}}$). Spatio-temporal visualizations (Fig. 2) confirm these trends, with CM exhibiting much higher levels of lateral (and rear) reflections in CM than in MML. These plots also confirm the perceived loudness difference between the two rooms, CM being judged louder.

In summary, similar trends were observed across systems for all attributes, with a weak effect (not significant) of system on *loudness*, *LEV*, and *readability*.

D. By musical extract

Results are presented by musical extract as shown in Fig. 6. Similar trends were observed between all attributes on both systems except for *reverberance* where binaural

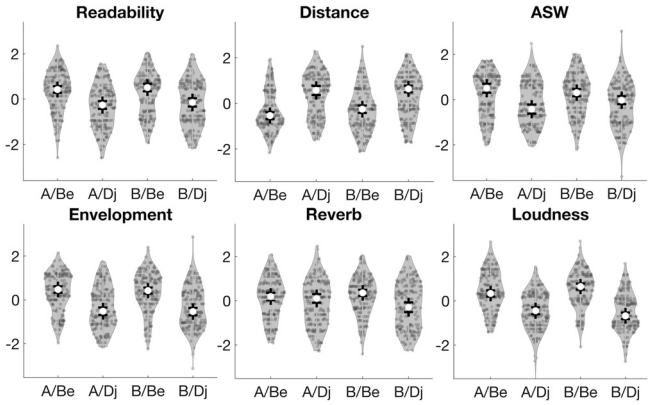

FIG. 6. Distribution of normalized results comparing (A)mbisonic and (B)inaural, for (Dj)ango and (Be)chet musical extracts. Results are aggregated over all subjects, positions, and rooms. (See Fig. 4 for plot details.)

https://doi.org/10.1121/10.0002942

headphone exhibited a significant difference due to extract while the loudspeaker condition did not (Bechet judged more reverberant than Django). It was observed with both systems that Bechet was judged more readable, closer, wider, more enveloping, more reverberant, and louder.

Regarding effect sizes, this independent variable was the most impacting one, providing the largest differences. The mean difference across all attributes is 0.3, with the largest being 0.77 (for *LEV*), emphasizing the influence of the musical extracts compared to room acoustics or even position in the room.

These results may be attributed to the actual content of the extract, Bechet being slower (80 vs 200 BPM), with a leading bright saxophone providing good readability, as compared to the background sustaining violin in Django. Bechet was also judged louder, due to the same saxophone and its frequency spectrum exciting the most sensible part of our hearing [between 1000 and 5000 Hz (Moore, 2012)]. Another potential explanation would be that the leading saxophone included dynamic directivity, providing intermittently more acoustic energy coming from the sides, known for increasing the spatial impression (i.e., *ASW* and *LEV* notably).

In summary, similar trends were observed for all attributes across systems, except for *reverberance* for which the sound reproduction system had a significant impact (although this could be attributed to the actual listening room).

E. Results summary

The results of the various tested conditions are summarized in Table V. It is examined if differences between acoustic configurations are consistently perceived across sound reproduction systems. If a significant difference between acoustic configurations is observed with one sound reproduction systems, but not observed in the other, then this indicates that the sound reproduction systems had a significant impact. Otherwise, it is suggested that perception of the given attributes is consistent across sound reproduction systems.

TABLE V. Summary of the results, showing the impact of the sound reproduction system. Each couple (attribute-independent variable) presents results by rendering system configuration (loudspeaker/headphones). Significant differences due to other conditions are represented by × and no significant differences are represented by –. The conditions for which the sound reproduction system has a significant impact (i.e., when there is an observed trend difference between system configurations) are indicated by ⊗.

Attr.	Position	Room	Musical extract
Readability	×/×	–/–	×/×
Distance	×/×	–/–	×/×
Loudness	×/×	×/×	×/×
ASW	–/⊗	×/×	×/×
LEV	⊗/–	×/×	×/×
Rev.	–/–	×/×	–/⊗

Overall, the sound reproduction system was observed to have an impact on *ASW, LEV*, and *reverberance* judgements: *ASW* and *LEV* differences were consistent across sound reproduction system configurations for room and extract changes, but not position changes. In contrast, *reverberance* differences were consistent for room and position changes, but not for musical extract.

When comparing front/back positions (across rooms and stimuli), although *ASW* was judged in both system conditions to be narrower at the front position than the back, this difference was only statistically significant in the binaural headphone condition, not in the loudspeaker condition.

Conversely, *LEV* was judged significantly higher at the front position in the loudspeaker condition, while the same trend was observed although not significant in the binaural headphone condition. In this instance, these effects appear to be interlinked, i.e., more enveloping being narrower. A much larger variance was also noted in responses for the back position/headphone listening condition, contributing to the lack of significance of the results.

When comparing rooms (across positions and stimuli), no significant differences were observed between sound reproduction system conditions, for all attributes. The same trends were observed equally, regardless of system.

When comparing stimuli (across positions and rooms), Django was judged significantly less *reverberant* than Bechet in binaural headphone, while judged comparably in the loudspeaker condition.

F. Interviews post-experiment

As mentioned in Gabrielsson and Sjogren (1979), it is beneficial to acquire various types of data to answer any research question, particularly when it involves humans in such multi-dimensional experiments. Therefore, short post-experiment interviews including the following questions were carried out, allowing participants to describe in their own words what they perceived:

- if they perceived any source movement (i.e., dynamic directivity), and for which source(s);
- which system they preferred, and the reasons why;
- if they perceived any differences between the two rendering systems;
- how they perceived each individual instrument: their level, position in the scene, and timbre.

In addition to these questions, they were also asked to rate on a 0 to 5 scale their skills or knowledge in the fields: *music playing, music listening, listening tests, and spatial audio*.

Results for preference showed an even split, with the loudspeaker system preferred by 7 out of 15 subjects and binaural headphone preferred by the remaining 8 subjects. Interestingly, the same reasons were given for both systems: the preferred system was perceived as more "immersive" and "more enveloping" for the majority of subjects (mentioned by 9 participants). Free-form comments also included

terms such as a *better sense of presence, pleasantness, clarity, distinctness, better localization, listening richness, details, and readability*. Regarding dynamic directivity/sense of source movement, it was not clearly perceived as such, but rather as a lively performance, non-static, or as timbral differences that provided a sense of movement (6 participants). In addition, this perception was only remarked for the saxophone, not the violin. This could be expected as the saxophone was louder and had a more leading role in a slower musical extract.

Additional interesting comments differentiating the two rendering systems, generally from the expert listeners, included: overall timbral difference between loudspeaker and binaural headphone systems, less externalization with headphones and a better sense of distance with loudspeakers. Untrained listeners often reported difficulties to rate some of the attributes (4 out of 6 participants), especially *ASW* and *LEV*, or even distinguishing two conditions. Regarding differences between stimuli, it was mentioned twice that the double-bass was sometimes too present, muddying the overall mix; some clearly identified the position of the sources (7 participants), while some perceived the scene shifted from the center in one direction or another (2 participants).

Finally, no significant effect was observed regarding a potential effect of the presentation order (participants starting with either of the system). No significant nor notable differences were observed between experienced and naive listeners when comparing musical extracts.

V. DISCUSSION

Subjective evaluations of second-order Ambisonic simulated auralizations of a jazz trio in several acoustic conditions have been assessed, comparing an 18 virtual-speaker binaural headphone to a 32 loudspeaker rendering. Similar variation trends were observed for all attributes with both rendering systems. Still, a significant effect of the sound reproduction system was observed for *ASW* and *LEV* when comparing front and back positions. This discrepancy may disappear or be reduced with a higher number of subjects (15 in this study). A significant effect of sound reproduction systems was also observed for *reverberance* (when comparing musical extracts), where differences were noted in one condition but appeared to be masked in the other. It is possible that the added contribution of the reverberation of the rendering room (albeit on the order of 0.2 s) was enough to mask the subtle differences perceived under headphone listening conditions.

As a result, the main hypothesis of auditory perception stability across sound reproduction systems must be rejected: the sound reproduction system had an observable impact, particularly on spatial attributes (*ASW, LEV*), and *reverberance*. As these attributes exhibited a large variance, more data would be needed to further investigate these differences, such as under more ideal rendering conditions (e.g., loudspeaker rendering in anechoic conditions, though this is an impractical situation for most installations).

Similar results were also observed for different VR visual rendering systems in the context of multi-modal auralizations (Thery *et al.*, 2017), evaluating mostly the same attributes (without *readability*), where *ASW* and *LEV* were also slightly impacted by the visual VR reproduction system. This could raise questions about the stability of these perceptual attributes in complex situations.

This relative stability can be put in perspective with previous comparisons where *consistent similarity* judgments were obtained between headphones and mono/multi-channel loudspeaker based rendering systems of audio only recorded material (Koehl *et al.*, 2011). However, as mentioned in the Introduction, this perceived *similarity* might be due to the too broad meaning of the attribute *similarity*, focusing predominantly on timbral attributes while potentially occluding specific spatial attributes.

This study also highlighted the strong dependence of acoustic attribute ratings on the stimuli content, as reported in Guastavino and Katz (2004). In the current study, two different musical extracts were presented, in various acoustic conditions (two rooms and two positions). A variance of room perception due to stimuli was observed, echoing early works on JNDs which showed a dependence of center time (i.e., clarity evaluations) and spatial impression to musical "motif" (Cox, 1993; Martelotta, 2010).

The fact that the *Django* musical extract was judged less reverberant than *Bechet* in binaural headphone, which was not the case in the loudspeaker condition where no significant difference appeared despite a similar trend. This difference may be due to the added acoustic response of the actual listening room, which provided an additional, though small, amount of reverberation to both stimuli (increasing the mean reverberation time of the simulation from 1.2 s by the 0.2 s reverberation time of the listening room to 1.4 s for the loudspeaker configuration. This slightly perceptible addition of reverberation could have smeared instruments' attack and release, especially for the faster *Django* extract, while *Bechet* could have been less affected due to the lower tempo (80 vs 200 BPM) with the leading bright saxophone also providing a better readability. Bechet was also judged louder, probably due to the same saxophone at certain moments in the extract being well above the mean level. The generally observed differences in *ASW* and *LEV* between musical extracts are likely due to spectral and directivity differences between the instruments. Dynamic directivity was included for the violin and saxophone, with the violin being a more "discrete" instrument, playing background harmonies, compared to the leading role of the saxophone. This movement could have intermittently provided more acoustic energy coming from the sides, more noticeable in the saxophone, thereby contributing to these spatial parameters.

A final point of interest in the results was the clear division in questionnaire responses. In terms of preference of rendering system, interviews resulted in an even split

https://doi.org/10.1121/10.0002942

between the two conditions, interestingly with the same aspects being mentioned to justify their choice in both cases. For their preferred rendering system, participants stated as potential factors that they felt more immersed, more enveloped, and that the instruments were clearer, terms that were already identified to contribute to sound reproduction system preference, both for experienced and inexperienced listeners (Francombe *et al.*, 2017). However, it is unclear how the same attribute arguments can be used to separate the systems, indicating either differences in how these attributes are interpreted (i.e., lack of formal listening training) or that the stated attributes are insufficiently precise to extract the meaningful information.

VI. CONCLUSION

This study has presented a subjective experiment in which participants were asked to rate the same second-order Ambisonic auralizations rendered over two systems: Head-tracked binaural decoding over an 18 virtual-speaker array presented over headphones, and direct decoding of the Ambisonic audio over a 32 loudspeaker 3D array in an acoustically damped room. The different auralizations represented two rooms of comparable size and reverberation time but of different form in order to highlight spatial variations in the acoustic response. Two receiver positions were defined in each room, maintaining proportional distances in each. Two musical extracts were selected from an anechoic Jazz trio recording database, providing two orchestrations and tempos for the listening test. Two instruments were modelled as fixed sources, while the center instrument was modelled to account for dynamic orientation of the musician, resulting in the inclusion of dynamic directivity during the extract.

Participants rated the different auralizations according to a set of six attributes commonly used in room acoustic evaluations: *readability, distance, apparent source width (ASW), listener envelopment (LEV), reverberance*, and *loudness*. Ratings were carried out using a fixed seven-point scale.

Overall, the observed differences and similarities in attribute ratings between room, position in the room, and musical extracts were consistent between Ambisonic binaural and loudspeaker renderings. However, a significant impact of the sound system was observed for *ASW* and *LEV* (when comparing positions in the room), though trends were similar. More importantly, a significant impact was observed for *reverberance* (when comparing stimuli), which is suggested as the effect of even a minor degree of room reverberation in the loudspeaker reproduction room masking differences observable over headphones. This suggests a recommendation of headphone over loudspeaker reproduction for detailed listening if ideal conditions are not achievable.

Auralization can thus be quite confidently used in acoustic design for decision making, although a wide range of stimuli should be provided to ensure the various uses of the designed space can be evaluated. This technology will probably benefit from the rapidly evolving market of virtual and augmented reality under the condition of sufficient training with the technology to ensure its adoption.

ACKNOWLEDGMENTS

The authors thank the managers of the Cité de la Musique for access to the hall and providing architectural plans.

Alary, B., Politis, A., Schlecht, S., and Valimaki, V. (**2019**). "Directional feedback delay network," J. Aud. Eng. Soc. **67**(10), 752–762.
Barron, M. (**2009**). *Auditorium Acoustics and Architectural Design*, 2nd ed. (Spon, London), p. 504.
Bech, S., Gulbol, M.-A., Martin, G., Ghani, J., and Ellermeier, W. (**2005**). "A listening test system for automotive audio part 2: Initial verification," in *118th Audio Engineering Society Convention*, pp. 1–19.
Carpentier, T., Noisternig, M., and Warusfel, O. (**2015**). "Twenty years of Ircam Spat: Looking back, looking forward," technical report.
Cohen, J. (**1988**). *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates, Mahwah, NJ).
Cohen, J. (**1990**). "Things I have learned (so far)," Am. Psychol. **45**(12), 1304–1312.
Cox, T. J., Davies, W. J., and Lam, Y. W. (**1993**). "The sensitivity of listeners to early sound field changes in auditoria," Acta Acustica united with Acustica **79**(1), 27–41.
Cumming, G. (**2014**). "The new statistics: Why and how," Psychol. Sci. **25**(1), 7–29.
Dalenbäck, B. I. (**2018**). "What is geometrical acoustics," technical report.
Engel, I., Lou Alon, D., Robinson, P. W., and Mehra, R. (**2019**). "The effect of generic headphone compensation on binaural renderings," in *73rd Audio Engineering Society Convention*, pp. 1–10.
Fischetti, A., Hemin, J., and Jouhaneau, J. (**1993**). "Differences between headphones and loudspeakers listening in spatial properties of sound perception," Appl. Acoust. **39**(4), 291–305.
Francombe, J., Brookes, T., Mason, R., and Woodcock, J. (**2017**). "Evaluation of spatial audio reproduction methods: Analysis of listener preference," J. Aud. Eng. Soc. **65**(3), 212–225.
Frank, M. (**2014**). "How to make ambisonics sound good," in *Forum Acusticum*, Krakow.
Gabrielsson, A., and Sjogren, H. (**1979**). "Perceived sound quality of sound-reproducing systems," J. Acoust. Soc. Am. **65**, 1019–1033.
Gerzon, M. A. (**1975**). "The design of precisely coincident microphone arrays for stereo and surround sound," in *50th Audio Engineering Society Convention*.
Guastavino, C., and Katz, B. (**2004**). "Perceptual evaluation of multi-dimensional spatial audio reproduction," J. Acoust. Soc. Am. **116**, 1105–1115.
Guastavino, C., Larcher, C., Catusseau, G., and Boussard, P. (**2007**). "Spatial audio quality evaluation: Comparing transaural, ambisonics, and stereo," in *International Conference on Auditory Display*, pp. 53–59.
ISO (**2009**). ISO-3382-1: "Acoustics—Measurement of room acoustic parameters—Part 1: Performance spaces" (International Organization for Standardization, Geneva, Switzerland).
ITU-R (**2015**). "Method for the subjective assessment of intermediate quality level of audio systems" (International Telecommunication Union, Geneva, Switzerland).
Jot, J.-M., Larcher, V., and Pernaux, J.-M. (**1999**). "A comparative study of 3-D audio encoding and rendering techniques," in *Audio Engineering Society Conference: Spatial Sound Reproduction*.
Katz, B., and Kahle, E. (**2008**). "Auditorium of the Morgan Library, computer aided design and post-construction results," in *Proceedings of IOA*, Vol. 30, pp. 1–8.
Koehl, V., Paquier, M., and Delikaris-Mania, S. (**2011**). "Comparison of subjective assessments obtained from listening tests through headphones and loudspeaker setyps," in *131st Audio Engineering Society Convention*, pp. 1–6.

Kowalski, M., Naruniec, J., and Daniluk, M. (**2015**). "Livescan3D: A fast and inexpensive 3D data acquisition system for multiple kinect v2 sensors," Int. Conf. 3D Vision **1**, 318–325.
Lokki, T. (**2011**). "Recording and reproducing concert hall acoustics for subjective evaluation," in *International Seminar on Virtual Acoustics*, pp. 32–37.
Majdak, P., and Noisternig, M. (**2015**). "AES standard for file exchange—Spatial acoustic data file format" (Audio Engineering Society, New York).
Marentakis, G., Zotter, F., and Frank, M. (**2014**). "Vector-base and ambisonic amplitude panning: A comparison using pop, classical, and contemporary spatial music," Acta Acust. Acust. **100**(5), 945–955.
Martelotta, F. (**2010**). "The just noticeable difference of center time and clarity index in large reverberant spaces," J. Acoust. Soc. Am. **128**(2), 654–663.
Moore, B. C. J. (**2012**). *An Introduction to the Psychology of Hearing* (Academic, New York), p. 457.
Nicol, R. (**2010**). "Représentation et perception des espaces auditifs virtuels," Habilitation à diriger des recherches, Université du Maine.
Noisternig, M. (**2003**). "A 3D ambisonic based binaural sound reproduction system," in *24th Audio Engineering Society Convention*, pp. 1–5.
Paquier, M., and Koehl, V. (**2015**). "Discriminability of the placement of supra-aural and circumaural headphones," Appl. Acoust. **93**, 130–139.
Pelzer, S., Maempel, H.-J., and Vorländer, M. (**2010**). "Room modeling for acoustic simulation and auralization tasks: Resolution of structural details," in *German Annual Conference on Acoustics (DAGA)*, pp. 709–710.
Poirier-Quinot, D., Postma, B., and Katz, B. F. G. (**2016**). "Augmented auralization: Complimenting auralizations with immersive virtual reality technologies," in *International Symposium on Music and Room Acoustics (ISMRA)*, Vol. 14, pp. 1–10.
Postma, B., Demontis, H., and Katz, B. (**2016**). "Subjective evaluation of dynamic voice directivity for auralizations," Acta Acustica Acust. **103**, 181–184.
Postma, B., and Katz, B. (**2016**). "Perceptive and objective evaluation of calibrated room acoustic simulation auralizations," J. Acoust. Soc. Am. **140**(6), 4326–4337.
Savioja, L., and Svensson, P. (**2015**). "Overview of geometrical room acoustic modeling techniques," J. Acoust. Soc. Am. **138**(2), 708–730.
Shabtai, N. R., Behler, G., Vorlander, M., and Weinzierl, S. (**2017**). "Generation and analysis of an acoustic radiation pattern database for forty-one musical instruments," J. Acoust. Soc. Am. **141**(4), 1246–1256.
Solvang, A. (**2008**). "Spectral impairment for two-dimensional higher order ambisonics," J. Aud. Eng. Soc. **56**(4), 267–279.
Stitt, P., Bertet, S., and Van Walstijn, M. (**2017**). "Off-center listening with third-order ambisonics: Dependence of perceived source direction on signal type," J. Aud. Eng. Soc. **65**(3), 188–197.
Tervo, S., Pätynen, J., Kuusinen, A., and Lokki, T. (**2013**). "Spatial decomposition method for room impulse responses," J. Aud. Eng. Soc. **61**(1/2), 17–28.
Thery, D. (**2020**). "Architectural auralizations: Towards the integration of virtual acoustic in architecture," Ph.D. thesis, LIMSI-CNRS, Université Paris-Saclay, Paris.
Thery, D., Boccara, V., and Katz, B. (**2019**). "Auralization use in acoustical design: A survey study of acoustical consultants," J. Acoust. Soc. Am. **145**(6), 3446–3456.
Thery, D., and Katz, B. (**2019**). "Anechoic audio and 3D-video content database of small ensemble performances for virtual concerts," in *International Congress on Acoustics*, pp. 739–746.
Thery, D., Poirier-Q, D., Postma, B., and Katz, B. (**2017**). "Impact of visual rendering system on subjective auralization assessment in VR," in *EuroVR* (Springer, Berlin), pp. 105–118.
Zacharov, N. (**2019**). *Sensory Evaluation of Sound* (CRC, Boca Raton, FL).
Zahorik, P. (**2002**). "Direct-to-reverberant energy ratio sensitivity," J. Acoust. Soc. Am. **112**(5), 2110–2117.
Zaunschirm, M., Schorkhuber, C., and Holdrich, R. (**2018**). "Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," J. Acoust. Soc. Am. **143**(6), 3616–3627.
Zotter, F., and Frank, M. (**2019**). *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality* (Springer, Berlin), p. 223.
Zotter, F., Frank, M., and Pomberger, H. (**2013**). "Comparison of energy-preserving and all-round ambisonic decoders," in *Annual German Conference on Acoustics (DAGA)*, pp. 1–4.