DOI: 10.61782/fa.2023.0605

forumacusticum 2023

LATENCY DETECTION THRESHOLD OF HEAD-TRACKING FOR DIFFERENT HEAD ROTATION SPEEDS IN BINAURAL RENDERING

Clément RAPPIN[1,2]* Julian PALACINO[1] Pascal RUEFF[1]
Laurent FEICHTER[1] Mathieu PAQUIER[2]
[1] Feichter Audio, 5 Rue Louis de Broglie, 22300 Lannion
[2] Université de Bretagne Occidentale, Lab-STICC (UMR 6285), 29200 Brest, France

ABSTRACT

When a head-tracking system is coupled with a binaural rendering, the latter becomes dynamic. Above all, a head-tracking system allows consistency between the orientation of the head and the audio rendering. This dynamism has many advantages, such as improved externalization, localization accuracy, or realism. However, an excessive latency between the head-tracking system and the audio rendering leads to a feeling of incoherence and/or slewing, which is detrimental to the listening experience.
In this paper, the detection threshold of head-tracking latency was estimated using an adaptive procedure. This task was performed using three head rotation speeds and a total of six different sound scenes : a pink noise, a male speech with and without reverberation, a pair of congas with and without reverberation, and a complex scene (a realistic coffee shop ambiance, containing six distinct sources with reverberation).

Keywords: *binaural, head-tracking, latency, head rotation speed, head movement*

1. INTRODUCTION

The techniques of binaural sound recording and rendering have been in full expansion for a few years, despite some limitations (individualization of Head-Related Impulse Response or HRIR, front/back confusions, spectral artifacts. . .). One of the main issues is the static nature of the rendering : if the listener moves his head, the virtual source follows his movement, whereas a real source would have remained motionless. A dynamic binaural rendering, taking into account the head movements, overcomes this issue. By using a head-tracking system, the listener can move naturally without altering the coherence between the orientation of his head and the relative position of the source. Moreover, the addition of a head-tracking system improves the quality of the audio rendering in several aspects : increase of the precision of localization, improvement of the feeling of externalization, reduction of the front/back confusions. . . [1–4]

However, the processing, transmission, and computation time of the data and the audio leads to an unavoidable latency. The Total System Latency, or TSL, can be defined as the delay between a head movement and the resulting sound output. If this latency is too high, a feeling of spatial slewing (or elasticity) can be felt on the sound source [5] or even lead to a strong spatial incoherence during a head movement.

Several studies have examined the latency detection threshold, in other words : at what delay a listener can perceive spatial slewing [6–9]. Different studies in the literature observed an average detection threshold between 50 ms and 110 ms for a single source according to different protocols. Different parameters were investigated : Stitt showed that this threshold increases with a complex scene [9], and Lindau showed that the excerpt or the reverberation has no significant impact [8]. In these studies, the head movements could be left free or imposed. No

*__*Corresponding author*__: clement.rappin@feichter-electronics.com.

statistical correlation was found between head movement, more precisely head rotation speed, and the latency detection threshold (Brungart [6] observed a trend, but without any statistical analysis). Nevertheless, the design of the tests did not take into account different speed conditions. When analyzed *a posteriori*, the differences in speed between subjects were moreover quite slight.

These elements motivated a new study on the latency detection threshold, involving different head rotation speeds. Several stimuli were also tested with and without reverberation.

2. METHOD

2.1 Stimuli

A total of 6 sound scenes of 8 s were presented in dynamic binaural rendering : a male speech, a pair of congas, a pink noise, and a coffee shop ambiance. The speech and congas scenes were presented with and without reverberation, while the pink noise was always presented without reverberation, and the coffee shop scene with reverberation. The sources remained stationary on the horizontal plane and were punctual. The precise position of each source can be visualized in Fig. 1.

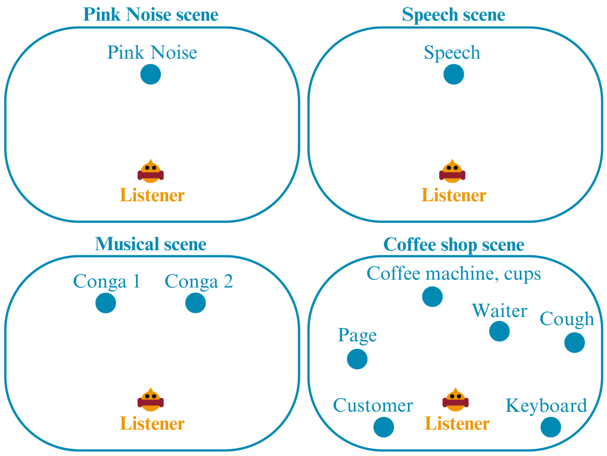


Figure 1: Sources positions for each audio scene.

2.2 Audio Rendering

The real-time audio processing (dynamic binaural) was carried out with Max/MSP. For each sound scene, the rendering engine was composed of two parts playing simultaneously. The first part was used to render the direct sound field in object-based format, while the second part only played the reverberation (if present in the scene) in ambisonic format. This separation allows the addition of reverberation while preserving an identical direct field, and is close to an ecological production mode.

The spatial audio workflow is presented in Fig. 2. The direct sound field was processed with the Spat5 tools [10]. Each object in the virtual scene corresponded to a sound source and was directly binauralized by convolution of the HRIR. The reverberation of all sound scenes was rendered with the convolution of the stimuli with SRIRs (Spatial Room Impulse Responses), in 4th order ambisonic format. SRIRs files were measured with an Eigenmike microphone from mhacoustics [11] with different positions of a loudspeaker. Each source was convolved by the SRIRs corresponding to its position in the virtual sound scene. The contribution of the direct sound field for each of the measurements was removed, leaving only the contribution of the reverberation.
The two parts of the audio engine (direct sound field and reverberation) were added to obtain the final sound output for each sound scene.

The listening level was fixed, between 65 and 80 dB SPL, depending on the excerpts. These differences in levels are explained by the natural difference in acoustic intensity between the stimuli (at an equal distance, a voice has a lower sound level than a pair of congas). Two elements are worth noticing : on one hand, for the stimuli presented with and without reverberation, the perceived level was equalized in loudness. On the other hand, the reverberation was treated as a mix effect and was not intended to be perfectly realistic (even if ecological validity was sought). It was therefore used to, among other things, create an effect of room and distance, especially during the complex scene.

The Neumann KU100 artificial head HRIR set (often employed in professional binaural productions), measured by Bernschultz [12], was used.

The minimal total system latency (TSL) was 31.73 ± 4.88 ms. For the test, this value varied by delaying the transmission of the head-tracking data to the audio renderer.

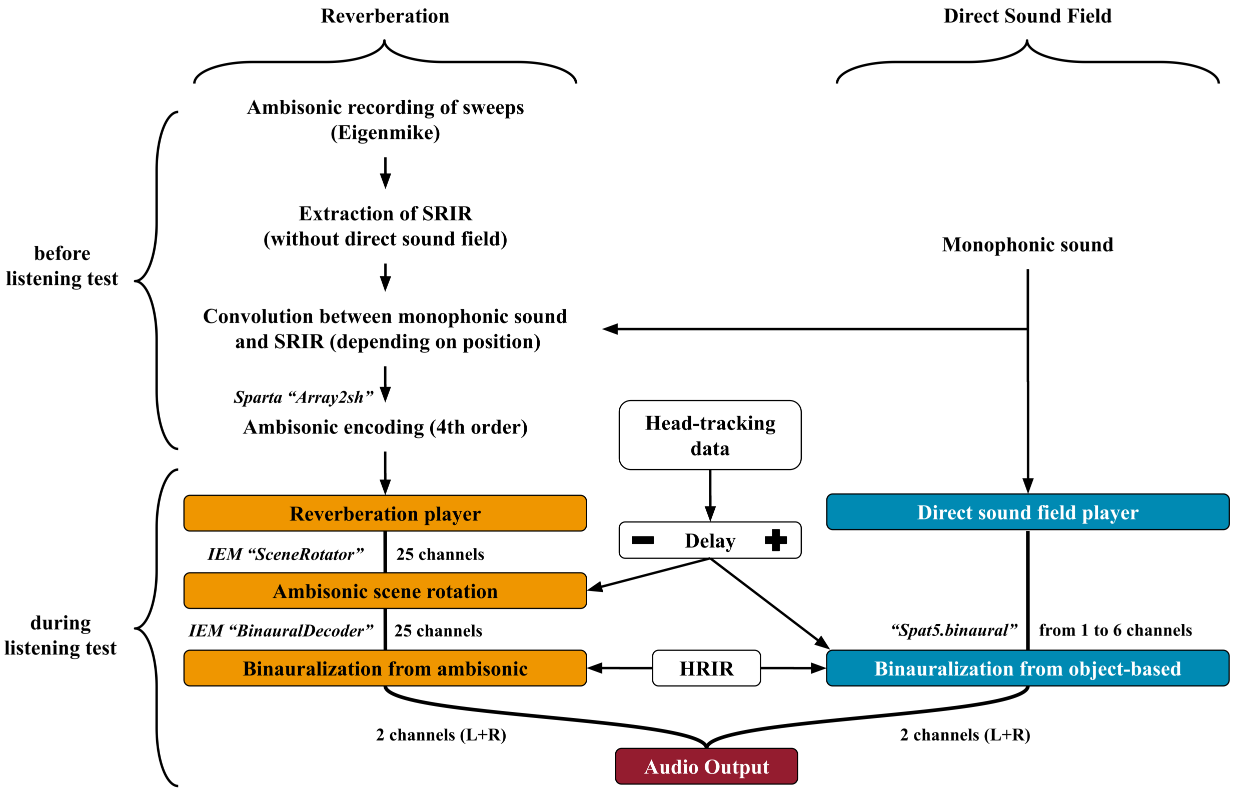


Figure 2: Audio Rendering synoptic [1] .

2.3 Protocol

2.3.1 Listening Test

Each subject was placed in a quiet recording studio environment. The stimuli were listened to with Sennheiser HD-650 headphones. A T3 (version 2) head-tracking system [13] measured the orientation of the head every 10 ms. On each trial, the subject listened to the stimulus and then answered the question, "Did you experience latency during your head movements ?" (translated from French : *"Avez-vous ressenti de la latence lors de vos mouvements de tête ?"*). The latency during the next excerpt was adjusted, following an adaptive procedure, close to the "1Up1Down" method. The initial latency was set to 182 ms (31.73 ms minimum latency + 150 ms additional latency), with an initial step of 30 ms. After 2 inversions, the step decreased to 20 ms, and after 4 new inversions to 10 ms. After a total of 10 inversions, it was assumed that sufficient data were obtained to calculate the latency detection threshold (equal to the average of the thresholds reached during the last 6 inversions), and another sound scene was played. The order of presentation of the sound scenes was randomized. The protocol and the interface were implemented in Max/MSP.
Before the test, the subjects were acquainted with the sensation of latency by an explicit definition and a sound example in which the latency was intentionally high (500 ms).

2.3.2 Head movements instructions

The test was divided into 3 sessions, corresponding to 3 different head movements (see Fig. 3) : a Slow Movement (SM, 60 °/s), a Medium Movement (MM, 180 °/s) and finally a Fast Movement (FM, 360 °/s). The subjects had to turn their head to the right, then straight ahead, then to the left, and finally straight ahead again, marking pauses between each movement (i.e., 4 pauses). In order for the

[1] More information can be found concerning IEM and SPARTA tools in [14–16].

subject to initiate these movements at the same moments as the stimulus, regardless of the condition, the duration of the pauses was dependent on the movement speed. We expected (from informal preliminary tests) that these pauses facilitate the detection of latency because sound sources keep moving after a sudden stop, or inversely.

As head movements were difficult to perform given the diversity of the sound scenes, subjects were trained at the beginning of each session. During the training phase, a visual cursor displayed on a large screen was representing the expected and the performed movement simultaneously. A metronome (sound click), which started playing a bar before the stimulus, was also used. During the test, there was no visual cursor and the metronome was played only before the stimulus (see Fig. 3). The experimenter was in the room during the training phase to check that the instructions were being followed correctly and to help the subject if necessary.

To ensure that movements were properly performed throughout the test, a small training phase went back after 6 inversions of the adaptive procedure, before a new excerpt, and if head movements were incorrect 3 times in a row.

All head-tracking data were observed and recorded. To avoid any tracker drift problems, the head-tracking system was calibrated before each trial.

2.4 Subjects

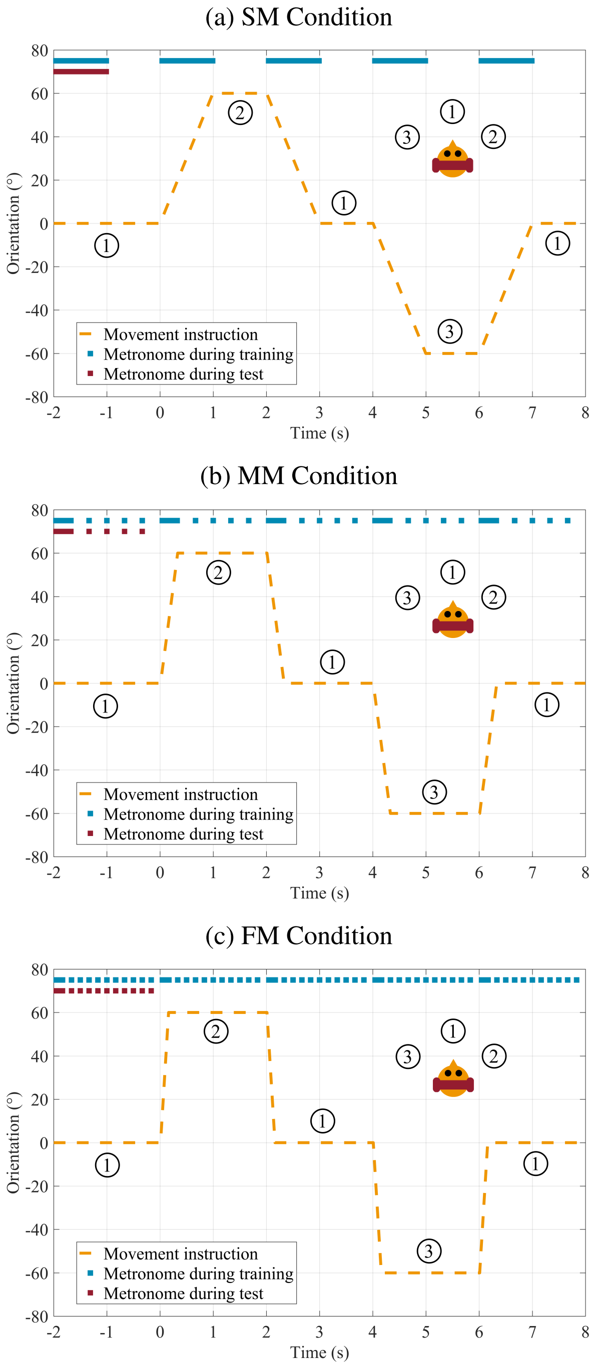
13 subjects participated in the study, aged between 20 and 26 years. Most of them were undergraduate students in a sound engineering training program. It was therefore assumed that the subjects had good critical listening skills, but were not especially experienced in binaural or spatial listening. None of them reported any hearing loss.

3. RESULTS

3.1 Head movement

Head movements were analyzed by calculating the average head rotation speed (pauses were excluded from the analysis). The results are reported in Fig. 4 for each movement instruction.

The subjects globally respected the movement instructions, despite average speeds lower than the target speeds. Nevertheless, subjects performed 3 head movements with significantly different speeds between them : $p < 0.001$ for each comparison with Wilcoxon test (speeds were not normally distributed).


(a) SM Condition

(b) MM Condition

(c) FM Condition

Figure 3: Movement instructions for all conditions. The stimulus started to play at time = 0 s. The metronome sound was continuous during head movement, and with distinct pulses during pauses

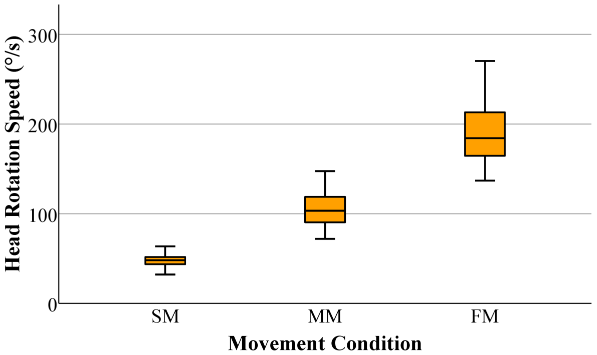


Figure 4: Mean head rotation speed across all subjects for each movement instruction : Slow Movement (SM), Medium Movement (MM), and Fast Movement (FM). The box represents Interquartile Range (IQR), the line in the box corresponds to the median and the whiskers to the values in 1.5 x IQR.

3.2 Latency Detection Threshold

Since the results were not normally distributed, a non-parametric analysis was used to study the effects of each condition. Fig. 5 represents the average detection thresholds obtained for each stimulus and each movement instruction. All presented thresholds include the minimum Total System Latency.

The median latency detection threshold across all subjects and all stimuli was 206.7 ms for the SM condition, 200.1 ms for the MM condition, and 195.0 ms for the FM condition. Despite small absolute differences, a Friedman test showed a significant effect of the movement speed on latency detection (p = 0.023). The Wilcoxon test revealed that the detection threshold for the SM condition was significantly higher than for the MM condition (p = 0.038) and the FM condition (p = 0.013). The MM and FM conditions were not significantly different (p = 0.864).

Regarding the influence of the stimulus, median latency detection thresholds ranged between 170.0 ms and 233.4 ms, for the pink noise and the coffee shop scenes respectively. A Friedman test showed a significant influence of the stimulus ($p < 0.001$). The detection threshold when listening to pink noise was significantly lower than for all the other stimuli ($p < 0.05$), except for the congas scene without reverberation. On the opposite, the detection threshold was significantly higher when listening to the coffee shop scene ($p < 0.05$). The detailed results are presented in the following Tab. 1

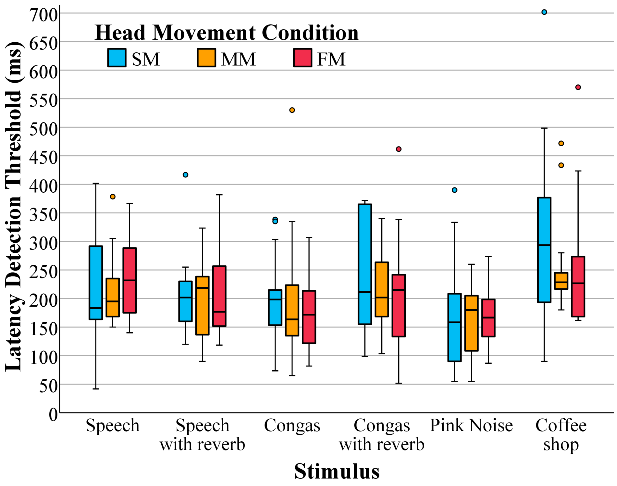


Figure 5: Latency detection threshold in relation to the stimuli for each movement instruction. The box represents Interquartile Range (IQR), the line in the box corresponds to the median and the whiskers to values in 1.5 x IQR.

Table 1: p-values of Wilcoxon test for each stimulus (averaged by head-rotation speed).

Stimulus	Speech	Speech (with rvb)	Congas	Congas (with rvb)	Pink Noise	Coffee shop
Speech		0.056	**0.010**	0.748	$<$ **0.001**	**0.012**
Speech (with rvb)	0.056		0.258	0.236	**0.016**	$<$ **0.001**
Congas	**0.010**	0.258		**0.004**	0.104	$<$ **0.001**
Congas (with rvb)	0.748	0.236	**0.004**		$<$ **0.001**	**0.016**
Pink Noise	$<$ **0.001**	**0.016**	0.104	$<$ **0.001**		$<$ **0.001**
Coffee shop	**0.012**	$<$ **0.001**	$<$ **0.001**	**0.016**	$<$ **0.001**	

4. DISCUSSION

Subjects had an overall lower threshold for fast movements, revealing a significant effect of head rotation speed on latency detection. The absence of a significant difference between the MM and FM conditions may indicate that there is a speed threshold for which detection is maximized.

The stimuli also had an influence on latency detection. Pink noise was the most critical excerpt for detection, in contrast to the coffee shop scene. The continuous nature of the stimulus seems to be a crucial factor in latency detection. Indeed, the complex scene was essentially composed of several sources playing one by one. It should also be noted that the coffee shop scene had a continuous source (coffee machine); some subjects verbally indicated that they were able to detect the presence of latency more easily by focusing only on this source. In addition, when listening to the speech scene, some subjects reported that they had more difficulty detecting latency because the pause time of the movement (the most critical moment for detection) could occur during a temporary stimulus interruption. This result is coherent with [9] who found that a more complex scene has a higher latency detection threshold.

Finally, conclusions about the influence of reverberation are difficult to draw. The results indicated that the addition of reverberation on the congas scene significantly increased the latency detection threshold, but the opposite trend was observed for the speech scene (with no significant effect, p = 0.056). The effect of reverberation depended on the excerpt, increasing, decreasing, or having no effect on the latency detection threshold. In a precedent study, Lindau [8] found that reverberation had no significant effect on latency detection.

The measured detection thresholds were higher than those observed in the literature. This difference could be explained by protocol differences, such as the use of non-individual HRIRs, subject expertise, or task. Performances of the most critical subjects were nevertheless similar to the literature, with latency detection thresholds as low as 44 ms across all excerpts and movements.

In the context of technological application, and regarding the large inter-individual differences, the most critical cases must be taken into account to ensure that a limited number of listeners will be likely to perceive latency. These most critical cases seem to be fast movements, with continuous stimulus. TSL lower than 40 ms seems to be fixed to ensure the inaudibility of latency.

5. CONCLUSION

The present study was conducted to measure the influence of head rotation speed on the latency detection threshold. In contrast to previous studies, a significant difference was perceived by the subjects between the fastest and slowest movement conditions. Significant effects were also observed between the different stimuli. A continuous sound source seems to be the most critical case, while a complex sound scene (composed of several sources) makes latency detection more difficult. The influence of the reverberation was not observed for all the excerpts. Large inter-individual differences were observed, which indicates that the performances between the subjects are strongly divergent.

6. REFERENCES

[1] D. R. Begault, E. M. Wenzel, and M. R. Anderson, “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source,” *Journal of the Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, 2001.

[2] S. Perrett and W. Noble, “The contribution of head motion cues to localization of low-pass noise,” *Perception & Psychophysics*, vol. 59, no. 7, pp. 1018–1026, 1997.

[3] E. Hendrickx, P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. F. Katz, and C. de Boishéraud, “Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis,” *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2011–2023, 2017.

[4] F. L. Wightman and D. J. Kistler, “Resolution of front–back ambiguity in spatial hearing by listener and source movement,” *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2841–2853, 1999.

[5] E. Wenzel, “Effect of increasing system latency on localization of virtual sounds,” in *Proc. of the AES 16th International Conference : Spatial Sounds Reproduction*, (Rovaniemi, Finland), 1999.

[6] D. S. Brungart, A. J. Kordik, and B. D. Simpson, “Effects of headtracker latency in virtual audio displays,” *Journal of the Audio Engineering Society*, vol. 54, no. 1, pp. 32–44, 2006.

[7] S. Yairi, Y. Iwaya, and Y. Suzuki, "Estimation of detection threshold of system latency of virtual auditory display," *Applied Acoustics*, vol. 68, no. 8, pp. 851–863, 2007.

[8] A. Lindau, "The perception of system latency in dynamic binaural synthesis," in *Proc. of the 35th DAGA*, (Rotterdam, Netherland), pp. 1063–1066, 2009.

[9] P. Stitt, E. Hendrickx, J.-C. Messonnier, and B. F. Katz, "The influence of head tracking latency on binaural rendering in simple and complex sound scenes," in *Proc. of the 140th AES Convention*, (Paris, France), 2009.

[10] "Spat," Last accessed April 2023. Ircam, https://forum.ircam.fr/projects/detail/spat/.

[11] mhacoustics, "Eigenmike microphone," Last accessed April 2023. https://mhacoustics.com/products.

[12] B. Benrschültz, "A spherical far field hrir/hrtf compilation of the neumann ku100," in *Proc. of the AIA-DAGA*, (Merano, Italy), pp. 592–595, 2013.

[13] "T3 head-tracking system," 2021 (Last accessed April 2023). Feichter Audio, http://feichter-audio.com/produits/diffusion/t3/.

[14] "Iem plug-in suite," Last accessed April 2023. https://plugins.iem.at/.

[15] L. McCormack, "Sparta - spatial audio real-time applications," Last accessed April 2023. https://leomccormack.github.io/sparta-site/docs/plugins/sparta-suite/.

[16] L. McCormack, S. Delikaris-Manias, A. Farina, D. Pinardi, and V. Pulkki, "Real-time conversion of sensor array signals into spherical harmonic signals with applications to spatially localised sub-band sound-field analysis," in *Proc. of the 144th AES Convention*, (Milan, Italy), 2018.