Zero-Delay Spatial Audio Rendering for Immersive Networked Music Performances

Christian Schörkhuber
atmoky GmbH
Graz, Austria
christian.schoerkhuber@atmoky.com

Markus Zaunschirm
atmoky GmbH
Graz, Austria
markus.zaunschirm@atmoky.com

Luca Turchet
Dep. of Information Engineering and Computer Science
University of Trento
Trento, Italy
luca.turchet@unitn.it

***Abstract*—Low-latency spatial audio processing is essential to reduce the overall latency in immersive networked music performances (NMP). In spatial audio rendering chains, most components, including source filtering for directivity, occlusion, and environmental effects, can be implemented with negligible latency using minimum-phase IIR filters. However, Head-Related Impulse Responses (HRIRs), which are fundamental to binaural rendering, often introduce non-negligible latency when measured HRIRs are used. In this paper, we discuss the sources of latency and present a zero-delay implementation based on minimum-phase FIR filters and delay lines. Through a controlled listening experiment involving 14 expert musicians, we evaluated the perceptual transparency of the zero-delay rendering with respect to a reference implementation without latency optimization. The experiment comprised two zero-delay versions with different filter lengths and involved samples for five different source directions and four different signal types. The results showed that participants could not distinguish the reference from a zero-delay version with 64 filter taps at 48 kHz sampling rate, indicating that this configuration is suitable for spatial audio rendering in NMP systems.**

***Index Terms*—Spatial Audio, Networked Immersive Audio, Musical Metaverse**

I. Introduction

The advent of high-speed, low-latency internet has fostered the development of Networked Music Performance (NMP) systems, enabling musicians to rehearse and perform together from remote locations [1]. For these interactions to be musically viable, achieving the lowest possible end-to-end latency is of utmost importance. End-to-end latencies above $30\,\mathrm{ms}$ are known to disrupt the tight temporal synchrony required for ensemble playing [2]. Beyond simply minimizing delay, creating a convincing and natural collaborative environment for musicians requires a sense of shared space. Spatial audio [3], [4] is crucial for establishing this immersive experience, providing the directional cues that musicians rely on to coordinate and interact [5]. However, the digital signal processing required to render spatial audio introduces its own latency.

The various sources of latency in immersive NMP systems, from network jitter and packetization to audio hardware buffers, have been recently identified and comprehensively discussed in [6]. While the delay from a spatial audio pipeline is often small compared to that of the network, it contributes to the overall "latency budget". Therefore, reducing the processing latency of the spatial audio pipeline is an important and tangible step towards minimizing the overall system latency.

The recent analysis reported in [7] has quantified the latency introduced by various spatial audio rendering tools, highlighting areas for potential optimization. This paper builds upon this previous work by examining the architecture of spatial audio rendering pipelines in more detail, with a specific focus on identifying components that are prone to introducing latency. Being the primary source of latency, we analyze the delay that is introduced when measured Head-Related Impulse Responses (HRIRs) are applied to the input signal. Based on this understanding, we implemented a zero-delay solution based on minimum-phase FIR filters and evaluated the perceptual transparency with respect to a reference implementation in a controlled listening experiment.

Our work aims at contributing to the field of networked immersive audio, which lately has attracted significant attention from the academic and industrial research community [8]–[13]. However, while methods to reduce latency have been explored with regard to networking [14], [15], less research has been conducted thus far on the reduction of latency introduced by the immersive audio rendering chain.

Notably, immersive NMP systems can be integrated in broader extended reality ecosystems, leading to more realistic forms of musical interactions over the network [16]–[18]. These results contribute to advancing the vision of a Metaverse dedicated to musical experiences [19].

II. Latency in the Spatial Audio Pipeline

y The creation of a plausible and shared acoustic environment is a crucial component for future NMP systems. A convincing simulation must account for a range of acoustic phenomena, including the directivity patterns of sound sources and various sound propagation effects such as acoustic diffraction, absorption, reflection, and late reverberation. As musicians in NMPs typically utilize headphones rather than extensive loudspeaker arrays, the system must also model the interaction of the simulated sound waves with the listener's head, torso, and ears. This is achieved by rendering the

This work has been supported by the MUSMET project funded by the EIC Pathfinder Open scheme of the European Commission (grant agreement n. 101184379).

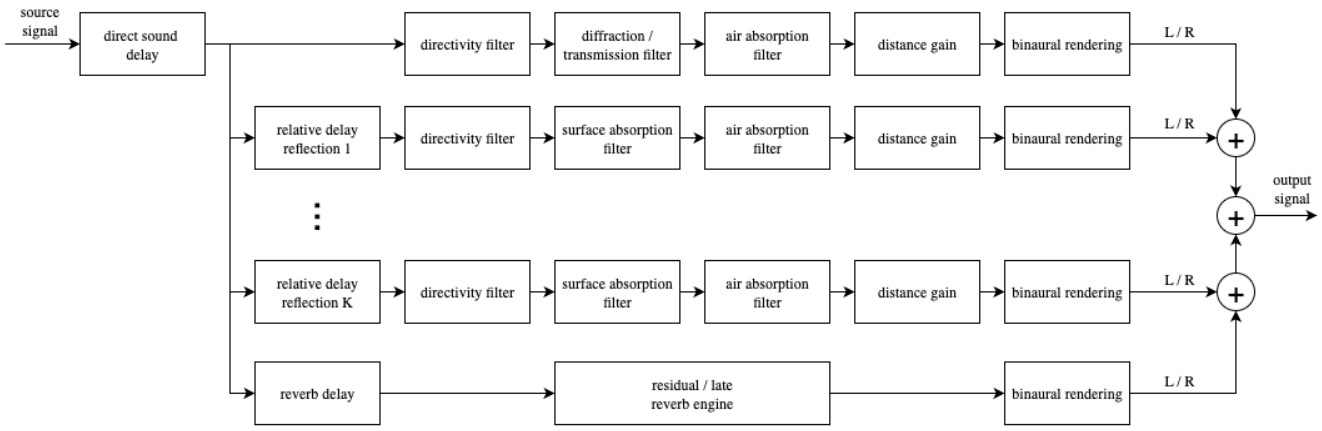

Fig. 1: Exemplary spatial audio processing pipeline for a single source.

complete sound field binaurally, often using head-tracking to update the rendering in real-time according to the listener's orientation.

That is, depending on the level of sophistication of the acoustic simulation, the raw signal of a source placed in a virtual space is subject to a series of processing steps before the output signal is played back. Figure 1 depicts a general overview of the processing blocks that such a simulation might contain for a single source. These include modeling the direct sound propagation with a series of filters and subsequent binaural rendering by applying Head-Related Transfer Functions (HRTFs) corresponding to the incident angle of the direct sound (which might not be the relative direction of the source if diffraction effects are considered), and similarly modeling K discrete early reflections as well as the early residual and the late reverb. All parameters for these processing steps, such as the position and orientations of the listener and the sound sources, as well as the propagation path for the direct sound and the reflections, are provided by a simulation engine. In the simplest case, such an engine only provides the relative position of the sources with respect to the listener, or it might update all sound propagation paths in real-time depending on changes of the acoustic environment. Nevertheless, the efficiency of the computation of the acoustic parameters, ranging from simple position updates to real-time sound propagation simulations, only affects the reaction latency of the rendering, not the processing latency, and is thus not relevant for the end-to-end latency of an NMP system. In contrast, the implementation of the signal processing steps outlined in Figure 1 has direct influence on the input-to-output latency.

In general, latency in digital signal processing algorithms stems from three primary sources: 1) look-ahead, 2) processing with windowed overlapping blocks, and 3) the group delay of linear filters. Look-ahead is common in dynamic processors like limiters and windowed block processing is commonly used in the context of time-frequency processing, e.g., for source separation, signal restoration, or adaptive beamforming. However, since the entire processing pipeline depicted in Figure 1 is a linear time-invariant (LTI) system, the only source of latency is the overall group delay and its latency can be simply measured by analyzing the impulse response of the overall system, where we define the latency of the system to correspond to the first significant peak. Furthermore, early reflections and late reverberation are, by their physical nature, delayed and filtered copies of the direct sound. Since they always occur after the direct sound arrives at the listener, they do not contribute to the latency of the system. Hence, only the direct path of the rendering pipeline is prone to introducing latency. We can group the processing blocks of the direct sound path as follows:

- **Direct sound delay:** This is an obvious source of latency, however, the time-of-flight from the source to the listener does not need to be modeled in the context of NMP, because it does not carry any perceptual information, as long as all reflections are delayed relative to the direct sound.
- **Source directivity filter:** The directivity of a real sound source is frequency-dependent and possibly time-varying, e.g. for musical instruments [20] or singing voice [21]. Applying linear-phase FIR filters obtained from interpolating measured source directivities [22] adds latency to the system which can be avoided, e.g., by approximating the measured directivity with IIR filter structures [23]. In the context NMPs, it often suffices to apply only a plausible, rather than an accurate simulation of source directivity. That is, latency can be avoided by modeling source directivity with a more generic, low-order minimum-phase IIR filter.
- **Propagation filters:** The same principle applies to filters modeling environmental effects, such as diffraction around obstacles, transmission through obstacles and air absorption. For the purposes of NMP, these effects can be approximated by low-order minimum-phase IIR filters, and hence do also not introduce any latency.

That is, for the purposes of NMP, where a plausible simulation

of the direct sound propagation from the source to the listener is sufficient - as opposed to accurate offline acoustic simulation - the zero-delay implementation of the filter stages via minimum-phase IIR filters is rather straightforward. Hence, in the remainder of the paper, we will discuss possible sources of latency in the last stage of the rendering pipeline, namely the binaural rendering stage.

III. Binaural Spatialization

The goal of binaural spatialization is to invoke the impression of a sound originating from a specific direction in three-dimensional space during playback over standard headphones. This is achieved by convolving a monophonic source signal with a pair of HRIRs corresponding to the desired direction. The HRIRs, often referred to by their frequency-domain equivalent, the HRTFs, encapsulate the complex filtering effects of the listener's anatomy - including the head, torso, and pinnae - on an incoming sound wave. These filters impose the critical acoustic cues for localization, namely the interaural time difference (ITD), the interaural level difference (ILD), and direction-dependent spectral coloration [3].

HRTFs are typically obtained through measurements, where impulse responses are captured from a loudspeaker positioned at various angles on a grid surrounding a human subject or a standardized dummy head with microphones placed at the ear canal entrances. Numerous such datasets are available in public databases (e.g., [24]–[26]). In a dynamic context like an NMP, where sound sources can move and listeners can change their head orientation, it is necessary to render arbitrary directions, including those not present on the original measurement grid. This requires a rendering method that is able to apply to the input signal an approximation of the true HRTF for any desired direction, given the HRTFs for the measured directions.

A. Rendering strategies

Similarly to the simulation of the directivity of the sound source, HRTFs can also be approximated by IIR filter structures, and several methods have been proposed to estimate filter coefficients given a measured HRTF set [23] [27]. However, while this is a very viable approach for efficient low-latency binaural rendering, perceptual transparency with respect to the original HRTFs using low-order filters is hard to achieve. Hence, in the remainder we will discuss different HRTF rendering methods based on FIR filters, which can be broadly categorized into three processing strategies.

- **Direct convolution:** When a new source direction is requested, a corresponding HRTF is computed via local interpolation considering the M closest points in the measurement grid, where M is typically 1 (nearest neighbor), 2 (linear interpolation), 3 (barycentric interpolation), or 4 (bilinear interpolation). To reduce the runtime complexity of this step, HRTFs can be precomputed for a regular spatial grid at initialization to allow for fast indexing and interpolation weight computation. Furthermore, to reduce spectral artifacts, the ITDs might be removed in the precomputed HRTFs and reapplied using a delay line.
- **Scene-based rendering:** The source signal is first encoded to an intermediate multichannel format, where the encoding - or panning - weights are recomputed whenever a new source direction is requested. A popular format in this context is Higher Order Ambisonics (HOA) [4], but other basis functions can also be used for this step [28]. The intermediate format is then rendered binaurally by applying a fixed filter matrix. If HOA is used, a popular filter design are MagLS-based approaches [29] as they achieve low perceptual HRTF approximation errors at relatively low Ambisonic orders. To incorporate head-rotation, a rotation matrix in the spherical harmonics domain is often applied prior to the application of the rendering filter.
- **Per-Ear scene-based rendering:** This is a variation of the scene-based approach, where two separate intermediate multichannel formats are used - one for each ear. In this case, the ITD is evaluated whenever a new direction is requested and applied to the input signal, before the signals are encoded and rendered for each ear. As the ITDs are handled explicitly, the filter matrices are constructed from a time aligned HRTF set, where the ITDs have been removed, which simplifies the filter design and potentially allows for a lower channel count of the intermediate format [30], [31].

Which of these rendering strategies is the most efficient heavily depends on the number of sources to be rendered. Scene-based renderer introduce a fixed performance cost due to the larger filter matrix, and the direct approach adds a convolution for each source. Hence, for a low number of sources, direct convolution is typically faster, and as the number of sources to be rendered increases, scene-based renders are more efficient. Different rendering strategies can also be used in parallel, such that the direct sound is rendered using direct convolution and all reflections are rendered using a scene-based approach (see e.g., [32]).

In any case, the choice of the specific rendering strategy does not have a direct implication on the latency: The goal of each of these methods is to approximate the true HRTF for any given direction as dictated by the chosen HRTF dataset, such that any of the above rendering methods is approximately equivalent to a convolution with the original HRTF. Hence, the latency that any of these rendering methods introduce is the latency that is inherent in the measurement data; more specifically, the group delays of the HRTFs.

B. Anatomy of HRTF latency

When HRTFs are measured, all sources have the same distance to the center of the head. Therefore, the time-of-arrival (TOA) of the source signal at the ears is direction dependent [34] due to the off-center position of the ears. That is, a source from 0° azimuth / 0° elevation reaches the left ear roughly $250\,\mu s$ (depending on the subject) later than a source from 90° azimuth / 0° elevation. Therefore, a "latency" of

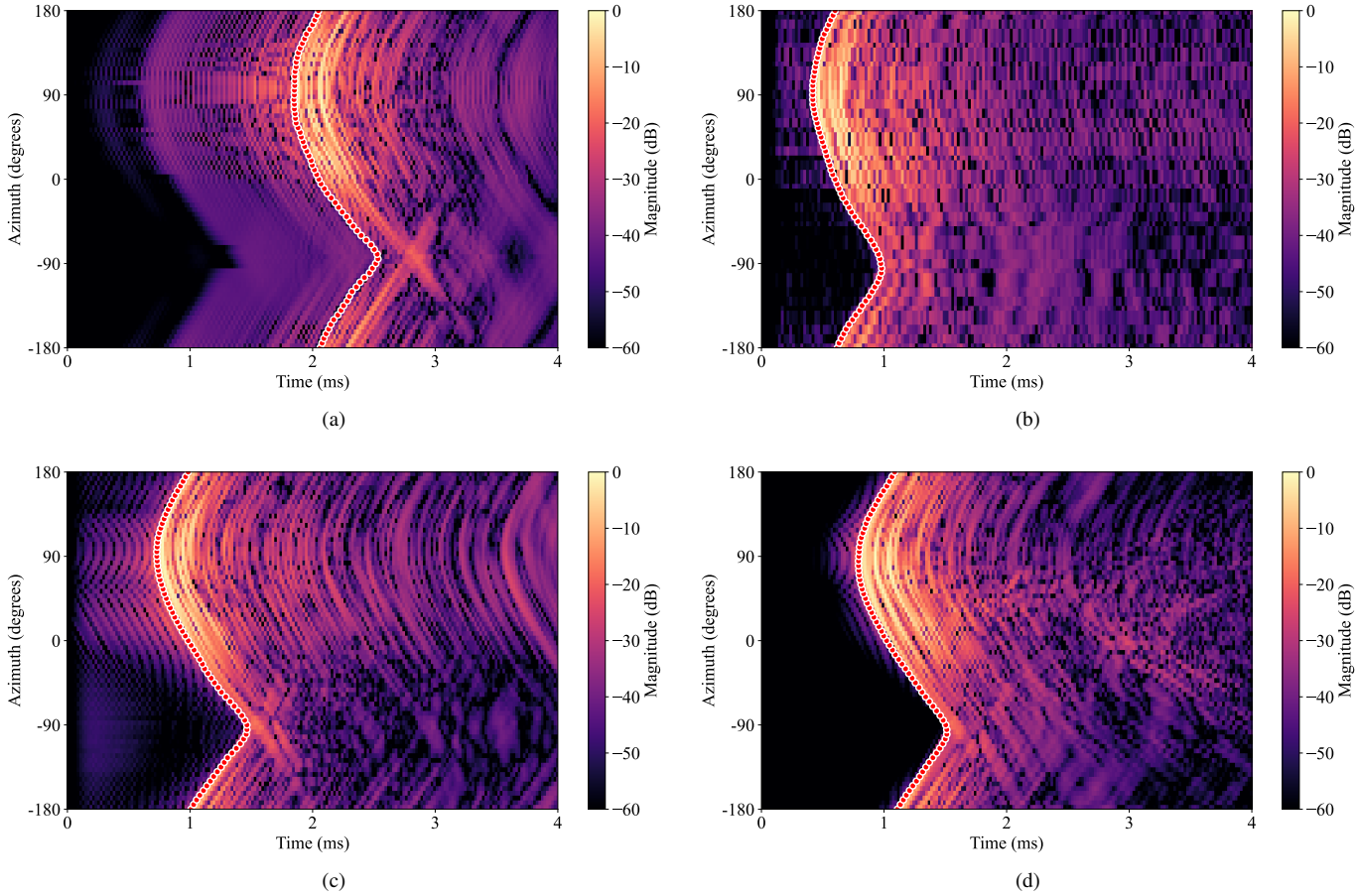

Fig. 2: Visualization of pre-ringing and time-of-arrival (TOA) variation in the left-ear horizontal HRIRs of four different public datasets. The red dots indicate the estimated TOA. (a) SADIEII (D1) [26], (b) HUTUBS (pp1) [24], (c) ARI (b nh172) [33] (d) SONICOM (P0001, minimum-phase free field compensation filter) [25].

around $250\,\mu s$ for frontal sources is physically induced. In theory, if the measurement was ideal, the measured HRTFs could be truncated so that the earliest TOA in the dataset is exactly 0, such that the only source of latency is due to the directional dependency of the TOA itself. However, the imperfections of the measurement equipment, specifically the loudspeakers and the microphones, introduce non-ideal magnitude and phase responses that are typically mitigated by applying equalization filters in a post-processing step. Additionally, the loudspeakers that are used typically cannot reproduce low-frequencies, and the low-frequency limit of the measurement chamber is often too high; therefore, the low-frequency responses of the HRTFs are modeled and merged with the original measurements using a cross-over filter (see e.g., [26]). Depending on the measurement equipment and the exact implementation of the compensation filters and the low-frequency extension, the resulting HRIRs include different amounts of pre-ringing, i.e., signal energy prior to the TOA of the first wave-front. If there is a significant amount of pre-ringing present in a specific HRIR, it cannot be simply truncated without introducing spectral artifacts. In Fig. 2 the HRIR magnitudes of the left ear for the horizontal plane are depicted for different datasets. In can be observed that different datasets exhibit different amounts of pre-ringing, whereas the SADEII [26] dataset has more prominent pre-ringing than, e.g., the HUTUBS [24] and the SONICOM [25] dataset. That is, the latency that is introduced by applying measured HRIRs varies between datasets, and depends both on the subject and the source direction. For example, the TOA of the ipsilateral ear for subject D1 in the SADIEII database varies between $1.85\,\mathrm{ms}$ and $2.1\,\mathrm{ms}$, and between $0.4\,\mathrm{ms}$ and $0.65\,\mathrm{ms}$ for subject pp1 of the HUTUBS dataset.

IV. Zero-delay implementation

In order to remove the latency of the binaural rendering stage, the measured HRTFs need to be pre-processed such that the TOA of the ipsilateral ear is always zero, while minimizing perceptual artifacts. Simply truncating the Head-Related Impulse Responses (HRIRs) before the TOA can introduce spectral artifacts which depend on the amount of pre-ringing. Therefore, we choose to convert all HRIRs in the dataset into a minimum-phase sequence using the cepstral

method [35]. For a given magnitude response, the minimum phase response has the smallest possible group delay such that the TOA for both ears is effectively set to zero, which means that the ITD needs to be added at a later stage. It is well known that HRTFs can be approximated with minimum-phase sequences and pure delays [36], [37], and while trained listeners may be able to detect the difference between the original HRTFs and their minimum-phase counterparts for lateral source directions [38], the perceptual ramifications are typically small. Another benefit of converting the HRIRs to a minimum-phase sequence is, that most of the signal energy is concentrated in the first part of the impulse response. That means that the resulting HRIRs can be truncated to reduce the computational complexity without introducing severe spectral artifacts. This is especially beneficial for NMP systems where the processing block size is rather small (e.g., 32 samples at sample frequency of $48\,\mathrm{kHz}$).

While in principle also scene-based rendering approaches could be implemented based on the minimum phase responses, we choose to adopt a direct convolution approach due to the relatively low number of expected direct sources. To this end, we precompute the minimum phase responses of the target HRIRs on a regular dense spatial grid by evaluating a HOA reference renderer designed according to [39].

At runtime, we fetch the minimum-phase HRIR for the desired direction via bilinear interpolation as described in section III-A. To achieve zero-latency for all directions, we always set the TOA of the ipsilateral ear to zero, and apply the full ITD to the contralateral ear as opposed to applying the physically correct direction dependent TOA. Since this is effectively only a constant time shift that affects both ears equally, this has no effect on the binaural cues for a specific direction. Note however, that a temporal shift of the HRIRs for a specific direction is only perceptually transparent when the source signals of different directions are uncorrelated, such that the relative time shift between directions is not perceivable. In case the signals from different directions are correlated, e.g., when they represent reflections of the same source signal, or when loudspeaker signals are virtualized, this assumption no longer holds. However, since the equivalent spatial shift is only about $8\,\mathrm{cm}$, the implications are negligible for NMPs, but might be worth considering for other use cases, e.g., for auditory research. The ITDs that are applied for each direction are estimated using the low-pass filtered thresholding method [40] and are spatially smoothed by fitting 5th order spherical harmonic coefficients. To avoid any offset latency due to fractional delay lines, we only use integer sample delays and apply a short fade when the ITD is updated.

The method has been implemented in C++ and can thus easily be integrated into an NMP system. To be able to conveniently evaluate the perceptual difference between the zero-delay implementation and the non-zero-delay reference, both methods have been wrapped as a VST plugin that allows for seamless switching between the two implementations. In Fig. 3 the measured impulse responses of the plugin in zero-delay mode are depicted for different source directions on the horizon.

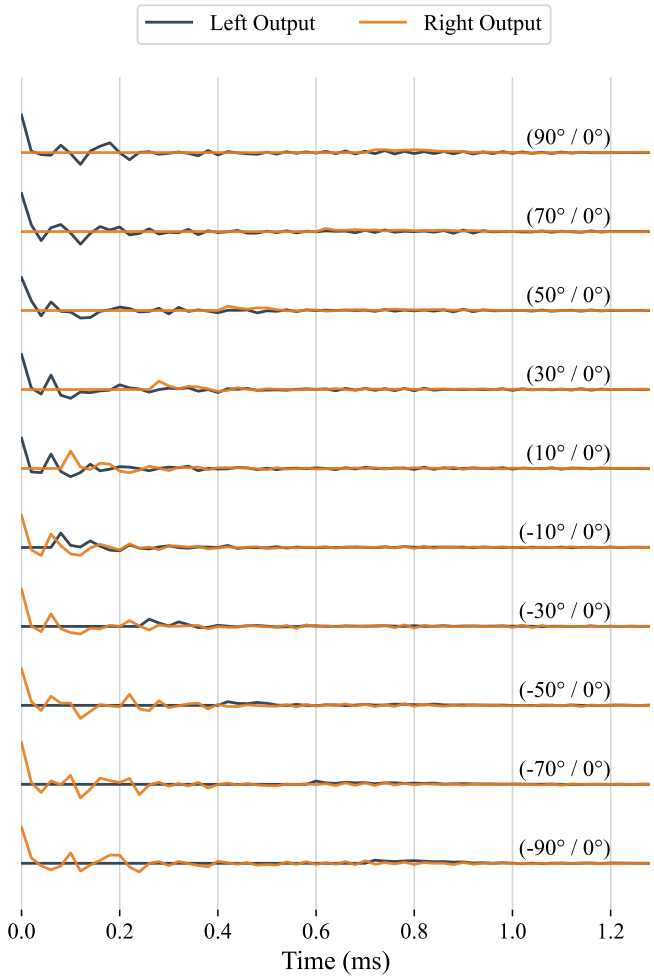

Fig. 3: Measured impulse responses of the spatializer plugin for different source directions on the horizontal plane.

V. Evaluation

A. Procedure and stimuli

To assess the efficacy of the implemented zero-delay plugin, we conducted a listening test. The test aimed to evaluate the perceived difference between the reference renderer and the zero-delay version of it, see IV. To ensure experimental control and repeatability, all test samples for the different conditions used renderings employing the VST plugin with pre-recorded source trajectories or directions.

The test comprised five different source direction conditions with the listener facing towards the front:

- *Static 1:* azimuth and elevation angles of (-35°, 0°);
- *Static 2:* azimuth and elevation angles of (-60°, 30°);
- *Static 3:* azimuth and elevation angles of (-145°, -25°);
- *Static 4:* azimuth and elevation angles of (80°, 0°);
- *Moving:* one continuous trajectory moving from the front clockwise to the back at 0° elevation with a constant speed of 22.5° per second at a constant distance.

The four lateral directions for the static sources were chosen due to the fact that the effects of the minimum-phase approximation and the application of a synthetic ITD are less noticeable for sources near the median plane.

For each source direction, four signal types were included in the test: speech, drums, piano, and white noise pulses. These four signal types were selected to cover a range of real-world signal classes and to be able to detect both spectral artifacts which are more prominent for broadband signals, and artifacts related to differences in group delay that are more prominent for transient signals.

The test also comprised four renderer versions (at a sampling rate of $48\,\mathrm{kHz}$):

- *Reference Renderer (Ref):* output of the non-zero-delay reference implementation ($L = 128$);
- *MP ($L = 64$):* output of the zero-delay implementation where the minimum-phase responses have been truncated to 64 samples;
- *MP ($L = 16$):* output of the zero-delay implementation where the minimum-phase responses have been truncated to 16 samples;
- *MP ($L = 8$, no ITD):* output of the zero-delay implementation where the minimum-phase responses have been truncated to 8 samples and no ITD has been applied. This version serves as anchor for the test.

Initial tests indicated that a truncation length of 64 samples introduces little to no audible artifacts compared to the full length minimum phase responses of 128 samples, hence this was chosen as the default configuration. The condition MP ($L = 16$) was included in the test to evaluate how much the perceived quality degrades when the minimum-phase responses are aggressively truncated.

As test procedure, we implemented the ITU-R BS.1534-3 MUlti Stimulus with Hidden Reference and Anchor (MUSHRA) test. Specifically, using webMUSHRA [41], we created a custom listening test by modifying the MUSHRA template.

In total, the test comprised 20 trials (4 signal types x 5 direction conditions). Each trial involved the evaluation the four renderer versions. The order of the trials was randomized across participants. Participants have been instructed to rate how well each sample matches the reference considering the perceived direction, the coloration, how well the sound is externalized, and the overall perceived similarity.

B. Participants

Fourteen participants (2 females, 12 males) aged between 26 and 44 (mean age = 35, SD = 6.11) took part in the listening test. All participants were musicians with at least 15 years of musical experience. All musicians used a pair of Beyerdynamic DT-770 Pro 80 Ohm headphones. Participants took on average 20 minutes to complete the experiment.

C. Results

An ANOVA was performed on a linear mixed effect model having the score, condition (reference, MP ($L = 64$), MP ($L = 16$), and anchor), signal (pulsed noise, speech, piano, and drums) and direction condition(4 static and 1 moving) as fixed factors, and the subject as a random factor. The assumption on the normality of the residuals was verified. Post hoc tests were performed on the fitted models using pairwise comparisons adjusted with the Tukey correction.

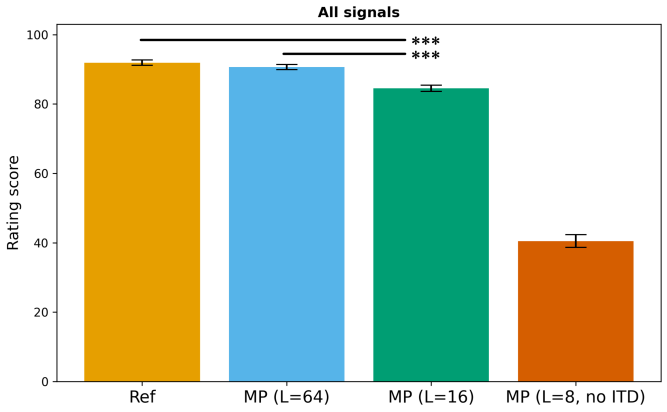

Fig. 4: Overall mean and standard error of the ratings for all types of signals. Legend: *** = p $<$ 0.001. Statistical significance is not displayed for pairs involving the anchor.

A significant main effect was found only for factor condition ($F(3, 1010.9) = 487.57$, $p < 0.001$, η_p^2 = 0.59, 95% CI [0.56,1]). In the following, as well as in Fig. 4, we report only the pairwise comparisons of interest for this study. No difference was found between reference and MP ($L = 64$). MP ($L = 16$) received significantly lower ratings than the reference (p $<$ 0.001, d = 0.4, 95% CI [0.22,0.56]) and MP ($L = 64$) (p $<$ 0.001, d = 0.34, 95% CI [0.16,0.49]). This may be attributed to the fact that i) timbral artifacts are introduced by the agressive filter truncation, especially at low frequencies, and ii) the ILD is changed by the truncation, because the impulse responses for the contralateral ear are typically longer than for the ipsilateral ear.

Fig. 5 graphically confirms that the trends above were shared for all the four types of signals involved in the study. Moreover, Fig. 6 shows that on average moving and static sound sources were consistently rated very similarly across all four types of signals.

VI. Discussion and Conclusion

This paper addressed the issue of latency optimization of spatial audio processing chains for integration into NMP systems. Firstly, this paper analyzed the sources of latency in HRIRs. Secondly, it presented a zero-delay implementation based on minimum-phase FIR filters.

Through a MUSHRA test involving 14 expert musicians, we evaluated the perceptual transparency of the implemented rendering with respect to a previous implementation that was not conceived for latency optimization. Beyond the reference and anchor, the test involved the conditions MP ($L = 64$) and MP ($L = 16$), which consisted of the output of the zero-

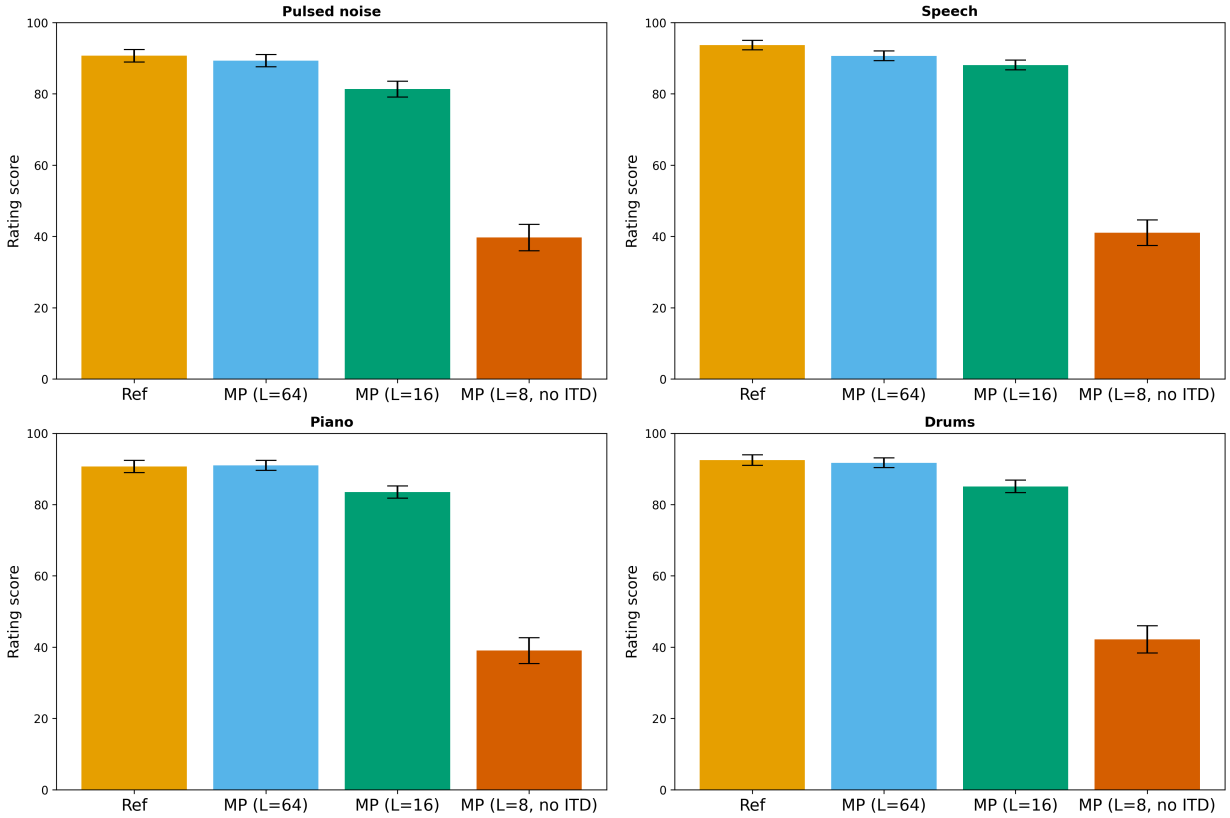

Fig. 5: Mean and standard error of the ratings for each type of signal.

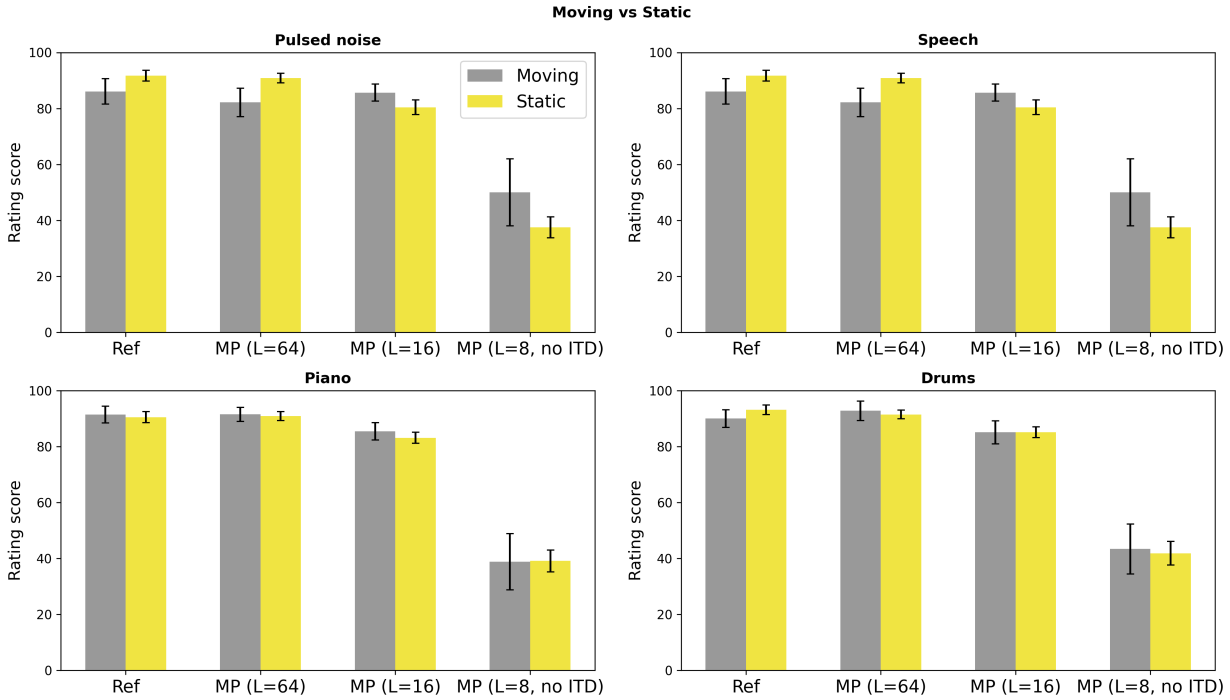

Fig. 6: Comparison between ratings of static and moving sources.

delay implementation where the minimum-phase responses were truncated respectively to 64 and 16 samples (at $48\,\mathrm{kHz}$).

Results showed that participants could not distinguish the original implementation from the zero-delay version of it in the MP ($L = 64$) condition, while they gave lower scores to the MP ($L = 16$) condition. This confirms that the zero-delay implementation using a 64-tap minimum-phase FIR filters can be used in place of the reference implementation without degrading the perceptual quality. Truncating the minimum-phase responses to only 16 samples, however, introduces noticeable spectral artifacts, mainly at low frequencies and for the contralateral ear, such that both timbral and ILD differences can be detected.

Based on our findings, in future works, we plan to integrate the developed method into an actual NMP system. We also plan to conduct objective measurements and subjective tests to technically and perceptually assess the validity of such integration.

ACKNOWLEDGMENT

We acknowledge the support of the MUSMET project funded by the EIC Pathfinder Open scheme of the European Commission (grant agreement n. 101184379). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Innovation Council. Neither the European Union nor the European Innovation Council can be held responsible for them.

REFERENCES

[1] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.

[2] K. Tsioutas and G. Xylomenos, "On the impact of audio characteristics to the quality of musicians' experience in network music performance," *J Audio Eng Soc*, vol. 69, no. 12, pp. 914–923, 2021.

[3] J. Paterson and H. Lee, *3D Audio*. Routledge, 2021.

[4] F. Zotter and M. Frank, *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer Nature, 2019.

[5] M. Tomasetti and L. Turchet, "Playing with others using headphones: musicians prefer binaural audio with head tracking over stereo," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 3, 2023.

[6] L. Turchet and M. Tomasetti, "Immersive networked music performance systems: identifying latency factors," in *Proceedings of the International Conference on Immersive and 3D Audio*, 2023.

[7] M. Tomasetti, A. Farina, and L. Turchet, "Latency of spatial audio plugins: a comparative study," in *Proceedings of the International Conference on Immersive and 3D Audio*, 2023, pp. 1–10.

[8] C. Rinaldi, F. Franchi, A. Marotta, F. Graziosi, and C. Centofanti, "On the exploitation of 5G multi-access edge computing for spatial audio in cultural heritage applications," *IEEE Access*, vol. 9, pp. 155 197–155 206, 2021.

[9] M. Multrus, S. Bruhn, J. Torres, E. Fotopoulou, T. Toftgård, E. Norvell, S. Döhla, Y. Gao, H. Su, L. Laaksonen *et al.*, "Immersive voice and audio services (IVAS) codec–the new 3GPP standard for immersive communication," in *Proc. of AES 157th Convention*, 2024.

[10] P. Cairns, H. Daffern, and G. Kearney, "Investigation of server–based spatial audio for metaverse concert distribution," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–8.

[11] J. Paulus, L. Laaksonen, T. Pihlajakuja, M.-V. Laitinen, J. Vilkamo, and A. Vasilache, "Metadata-assisted spatial audio (MASA)–an overview," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–10.

[12] S. Giacomelli, C. Centofanti, J. Santos, M. Galbiati, T. Salvi, F. Graziosi, and C. Rinaldi, "Remote immersive audio production: State of the art implementation, challenges, and improvements," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–10.

[13] C. Rinaldi and C. Centofanti, "The musical metaverse: Advancements and applications in networked immersive audio," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–7.

[14] F. Martusciello, C. Centofanti, C. Rinaldi, and A. Marotta, "Edge-enabled spatial audio service: Implementation and performance analysis on a MEC 5G infrastructure," in *2023 4th International Symposium on the Internet of Sounds*. IEEE, 2023, pp. 1–8.

[15] L. Turchet, C. Rinaldi, C. Centofanti, L. Vignati, and C. Rottondi, "5G-enabled internet of musical things architectures for remote immersive musical practices," *IEEE Open Journal of the Communications Society*, 2024.

[16] A. Boem, M. Tomasetti, and L. Turchet, "Issues and challenges in audio technologies for the musical metaverse," *Journal of the Audio Engineering Society*, vol. 73, pp. 94–114, 2025.

[17] A. F. Genovese, M. Gospodarek, Z. Nguyen, R. Pahle, and A. Roginska, "Locally adapted immersive environments for distributed music performances in mixed reality," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–10.

[18] A. F. Genovese, Z. Nguyen, M. Gospodarek, R. Pahle, C. Brenner, and A. Roginska, "Holodeck: A research framework for distributed multimedia concert performances," in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–10.

[19] L. Turchet, "Musical Metaverse: vision, opportunities, and challenges," *Personal and Ubiquitous Computing*, pp. 1–17, 2023.

[20] D. Ackermann, F. Brinkmann, and S. Weinzierl, "A database with directivities of musical instruments," *Journal of the Audio Engineering Society*, vol. 72, no. 3, pp. 170–179, 2024.

[21] M. Brandner, R. Blandin, M. Frank, and A. Sontacchi, "A pilot study on the influence of mouth configuration and torso on singing voice directivity," *The Journal of the Acoustical Society of America*, vol. 148, no. 3, pp. 1169–1180, 2020.

[22] D. Ackermann, F. Brinkmann, F. Zotter, M. Kob, and S. Weinzierl, "Comparative evaluation of interpolation methods for the directivity of musical instruments," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, p. 36, 2021.

[23] S. D. Ewert, O. Buttler, and H. Hu, "Computationally efficient parametric filter approximations for sound-source directivity and head-related impulse responses," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, 2021, pp. 1–6.

[24] B. Fabian, D. Manoj, P. Robert, W. Jan Joschka, S. Fabian, V. Daniel, G. Peter, and W. Stefan, "The HUTUBS head-related transfer function (HRTF) database," 2019. [Online]. Available: http://dx.doi.org/10.14279/depositonce-8487

[25] I. Engel, R. Daugintis, T. Vicente, A. O. Hogg, J. Pauwels, A. J. Tournier, and L. Picinali, "The SONICOM HRTF dataset," *journal of the audio engineering society*, vol. 71, no. 5, pp. 241–253, 2023.

[26] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database," *Applied Sciences*, vol. 8, no. 11, p. 2029, 2018.

[27] P. Nowak, "Spatial audio through headphones based on HRTFs approximated by parametric IIR filters," Ph.D. dissertation, Universitätsbibliothek der HSU/UniBw H, 2022.

[28] M. Marchan and A. Allen, "Multi-layered architecture for efficient and accurate HRTF rendering," *Journal of the Audio Engineering Society*, vol. 71, no. 6, pp. 338–348, 2023.

[29] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural rendering of Ambisonic signals via Magnitude Least Squares," in *Proceedings of the DAGA*, vol. 44, 2018, pp. 339–342.

[30] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Binaural reproduction based on bilateral ambisonics and ear-aligned HRTFs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 901–913, 2021.

[31] J.-G. Richter, M. Pollow, F. Wefers, and J. Fels, "Spherical harmonics based HRTF datasets: Implementation and evaluation for real-time auralization," *Acta Acustica united with Acustica*, vol. 100, no. 4, pp. 667–675, 2014.

[32] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona,

"3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation," *PloS one*, vol. 14, no. 3, p. e0211899, 2019.

[33] P. Majdak, B. Masiero, and J. Fels, "Sound localization in individualized and non-individualized crosstalk cancellation systems," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2055–2068, 2013.

[34] H. Ziegelwanger and P. Majdak, "Modeling the direction-continuous time-of-arrival in head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1278–1293, 2014.

[35] A. V. Oppenheim, *Discrete-time signal processing*. Pearson Education India, 1999.

[36] A. Kulkarni, S. Isabelle, and H. Colburn, "Sensitivity of human subjects to head-related transfer-function phase spectra," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2821–2840, 1999.

[37] J. Nam, M. A. Kolar, and J. S. Abel, "On the minimum-phase nature of head-related transfer functions," in *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[38] A. Andreopoulou and B. F. Katz, "Comparing the effect of HRTF processing techniques on perceptual quality ratings," in *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[39] C. Schörkhuber, "Method for generating a conversion filter for converting a multidimensional output audio signal into a two-dimensional listening audio signal," Patent AT 523 644 B1, Jun. 15, 2022.

[40] A. Andreopoulou and B. F. Katz, "Identification of perceptually relevant methods of inter-aural time difference estimation," *The Journal of the Acoustical Society of America*, vol. 142, no. 2, pp. 588–598, 2017.

[41] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA—A comprehensive framework for web-based listening tests," *Journal of Open Research Software*, 2018.