Edge-Enabled Spatial Audio Service: Implementation and Performance Analysis on a MEC 5G Infrastructure

Federico Martusciello[1], Carlo Centofanti[2], Claudia Rinaldi[3], and Andrea Marotta[2]

[1]Independent
[2]DISIM - University of L'Aquila, L'Aquila, Italy
[3]CNIT - Consorzio Nazionale Interuniversitario per le Telecomunicazioni, L'Aquila, Italy

***Abstract*—Spatial audio technologies are becoming a fundamental requirement for guaranteeing immersive auditory experiences in various applications such as Augmented and Virtual Reality up to the Metaverse. With the rise of mobile and edge computing, there is a growing interest in exploring spatial audio algorithms performance on edge infrastructures. This paper presents an evaluation of two different spatial audio algorithms and the potential for offloading the real time spatial audio processing on a Mobile Edge Computing (MEC) infrastructure. The presented results were obtained through evaluations performed on a real operator network, and they demonstrate the feasibility of offloading the computation of spatial audio algorithms at the network edge. No difference in terms of performance between the two algorithms was observed under the assumed scenario.**

***Index Terms*—Metaverse, immersivity, spatial audio, binaural rendering, multi-access edge computing, 5G, networked music performance**

I. Introduction

Recent advances in communication technologies as 5th Generation (5G) and beyond have pushed the development of new applications, moving from public safety [1], [2], up to networked music performance [3], virtual reality [4], and augmented reality [5], whose final arrival seems to be, so far, the metaverse [6]. A unified definition of the concept of metaverse does not exist so far, it is surely a concept popularized by science fiction literature and films, which has rapidly transformed from a futuristic vision to an imminent reality. Basing on [7], the characteristics of the metaverse are:

- immersivity: it has to guarantee a realistic experience to the users;
- persistence: it has to exist even if none is logged in;
- interactivity: it has to guarantee a high level of interaction between users through a large variety of sensors;
- socialization: it has to allow the users to share their experiences and knowledge trough various platforms;
- customization: the virtual environment must be personalised;
- economy: it has to include a virtual economic system for commercialization of services and virtual objects.

In this paper we focus on a solution to increase immersivity in the metaverse through spatial audio techniques for auralization [8]. The latter term is referred to the process of creating an auditory representation or simulation of a physical space or an acoustic environment. It allows a user to listen to how a sound would propagate and be perceived in a given space. Various solutions exist in the literature to achieve this purpose, some of them are proprietary as Dolby ATMOS, Sony 360 Reality Audio, Sennheiser AMBEO, while others come from open scientific research such as Vector Based Amplitude Panning (VBAP) [9], Distance Based Amplitude Panning (DBAP) [10], Ambisonics [11]. These solutions are usually employed on multichannel loudspeakers setup, instead, when the aim is to recreate the perception of sound through headphones, binaural rendering has to be used [12]. It takes into account the filtering and modifications that occur as sound reaches each ear, including interaural time differences (ITDs) and interaural level differences (ILDs). Binaural rendering relies on Head-Related Impulse Responses (HRIRs) to capture the individualized acoustic characteristics of the listener and provide an accurate spatial audio experience for headphones playback [13].

Since the concept of metaverse is strictly related to the Internet, dealing with spatial audio on the metaverse implies relating to the emerging paradigm of Internet of Sounds (IoS), given that the considered scenario refers to audio processing in a networked context [14]. The requirements imposed by this scenario can be satisfied by relying on a efficient and reliable communication link. In this context 5G and beyond communication networks play a fundamental role. 5G synergy with Virtual Network Functions (VNFs), Multi-Access Edge Computing (MEC), and Network Slicing forms the backbone of a dynamic and efficient network ecosystem, where services can be virtualized, localized, and customized to increased flexibility at the application layer [15]. These capabilities may

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART") and by the Italian Government under CIPE resolution no. 135 (December 21, 2012), project INnovating City Planning through Information and Communication Technologies (INCIPICT).

in turn allow to address some of the IoS challenges such as (i) platform independence by only sending uncompressed audio signals to the final users, (ii) low latency audio by exploiting important computation capabilities at the network side (iii) optimization of resources consumption by lightening the computational burden at the user side.

In this paper, we want to evaluate the effects of moving the computation of the spatial audio service to the network, as close as possible to the user, through the MEC paradigm. To the authors knowledge, this is one of the first papers dealing with spatial audio over MEC that presents an effective implementation of a headtracked binaural service on a real 5G network. This procedure is expected to show the double advantage of lightening the computational weight at the local user, by offloading the most of spatial audio service computations to the network side, while maintaining the latency very low due to the physical closeness of the MEC server to the final user, thus keep on guaranteeing a high Quality of Experience (QoE) as a local implementation.

The remaining of the paper is organized as follows: Section II describes the context of the presented study, discussing the exploited binaural rendering algorithms in II-A, while network architectures enabling the feasibility of the proposed scenario are discussed in II-B. Section III presents the technical solutions adopted for the implementation of the case study, while results are discussed in Section IV. Finally conclusions and future works are drawn in Section V.

II. Scenario

In this preliminary study, we considered applications with only three degrees of freedom (3DoF), thus rendering the virtual sound with respect to head movements along three possible axes (i.e. yaw, pitch, roll), even if the plan is to implement a whole AR service at the edge allowing the user to also move in the virtual space along x, y, z coordinates (six degrees of freedom [6DoF] [16]).

The scenario is composed by a final user, wearing headphones, who is listening to a sound source in a virtual environment. The sound may stand still or move along a trajectory which is independent from the user's movements, and since the whole virtual soundfield must be maintained, the two audio channels at the left and right ears of the listeners are computed through binaural rendering solutions that take into account head movements. The binaural rendering computation is offloaded to the server, while the client, i.e. the final user, transmits head positions data and receives two audio channels. In particular we exploit two different methods to obtain binaural audio, i.e. the classical one based on convolutions with changing HRIRs and another based on the concept of virtual loudspeakers in Ambisonics as explained below.

A. Binaural rendering algorithms

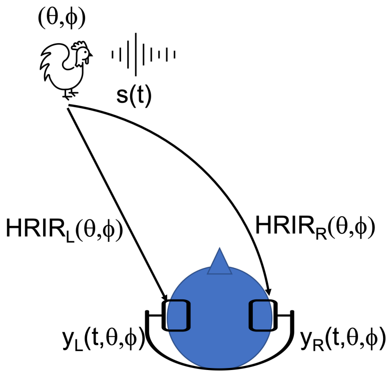


Fig. 1. Simple binaural rendering.

Despite the important advantages introduced by next generation communication protocols, the physical distance between elements of the network is a source of packet losses and delays that has to be considered while implementing a remote immersive service.

It is well known that the minimum threshold for the detection of latency in dynamic binaural listening can be fixed at 50 ms, based on perception studies as [17]. Thus, given the inherent network limitations, it is important to properly chose the spatialization algorithm and binaural rendering solutions to reduce as much as possible the computational latency.

In section I we have already discussed about the large availability of sound spatialization solutions. In this section we detail the two solutions that have been considered to make tests on the network infrastructure. As a baseline solution, the typical binaural rendering obtained by convolving the virtual sound source position with the corresponding HRIRs has been assumed, see Fig. 1, [18]. That is, considering the mono sound source $s(t)$ to be located at (θ, ϕ) with respect to the listener position, the signals to be reproduced to channels left (L) and right (R) of the headphones are:

$$y_L(t, \theta, \phi) = s(t) \otimes \mathrm{HRIR}_L(\theta, \phi) \tag{1}$$

$$y_R(t, \theta, \phi) = s(t) \otimes \mathrm{HRIR}_R(\theta, \phi) \tag{2}$$

The second solution for the implementation of binaural rendering is based on the idea of virtual loudspeakers [19]. It basically assumes to virtually decode the spatialized signal on a certain number of loudspeakers, which, for binaural rendering purposes, are in turn considered as virtual sound sources and thus treated following the baseline procedure previously discussed, this is reported in Fig. 2. This approach has the great advantage of maintaining fixed the position of the sound sources (i.e. the positions of the virtual loudspeakers), because the head rotation is converted into an opposite rotation of the whole soundfield before the decoding phase, as described in Fig. 3.

B. Network architecture

The 3GPP (3rd Generation Partnership Project) defines a number of deployment options to ease the transition from 4G to 5G standard. The choice on how to deploy 5G is up to the telecommunication operators and depends by many factors including licensed spectrum, geographical and morphological

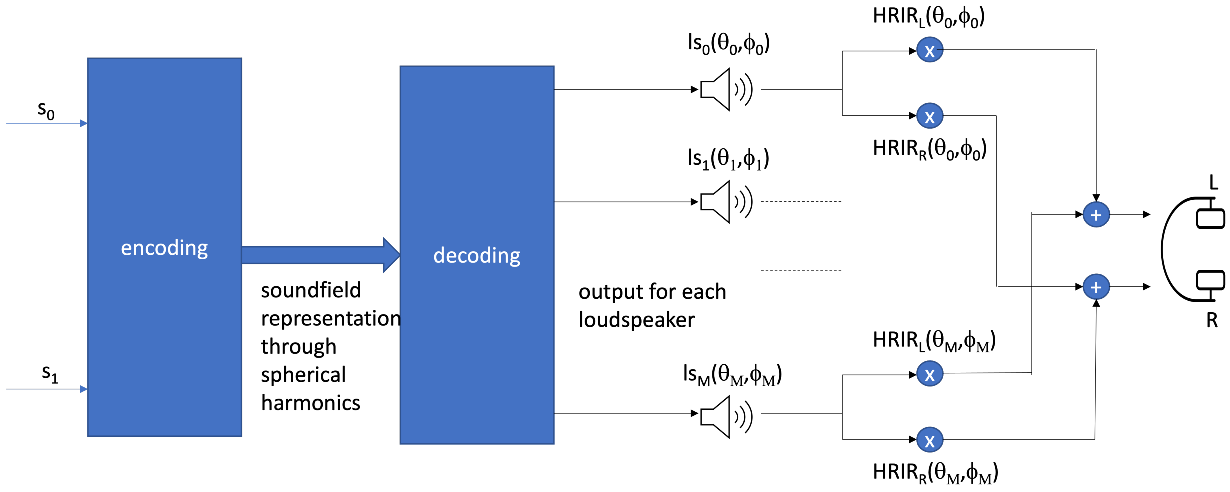

s_0
s_1
encoding
soundfield representation through spherical harmonics
decoding
output for each loudspeaker
$ls_0(\theta_0,\phi_0)$
$ls_1(\theta_1,\phi_1)$
$ls_M(\theta_M,\phi_M)$
$HRIR_L(\theta_0,\phi_0)$
$HRIR_R(\theta_0,\phi_0)$
$HRIR_L(\theta_M,\phi_M)$
$HRIR_R(\theta_M,\phi_M)$
x
+
L
R

Fig. 2. Virtual loudspeakers based binaural rendering.

aspects, capabilities of chosen equipment, and business factors. The current mobile networks rely on GTP (GPRS Tunnelling Protocol) to encapsulate data-plane traffic into layer 2 frames. The main purpose of encapsulating packets into GTP protocol is to support user mobility which is something that the IP protocol itself does not provide by default. This means the user traffic has no visibility of the IP layer 3 so not being able to forward packets to the final destination prior being decapsulated from the 5G core.

The network architecture considered is a Non Stand Alone (NSA) 5G deployment as per the telco operator's availability with whom we collaborate. It is worth noticing that telco operators have not yet deployed a 5G SA architecture so far in Italy, so for a real network implementation we had no choice.

Fig. 4 depicts the NSA deployment. Focusing on the bottom part of the figure it is possible to isolate a subgroup of the deployment composed by UE2, UE3, 5G Edge Core, and Multi-access Edge Computing that itself can be considered as a Stand Alone (SA) deployment for the application that can be run locally, without requiring the connectivity to a cloud server. The NSA architecture shown on the upper part of Fig. 4 is characterised by the need of reaching the Data Center Facilities layer to enable communication from different users coming from the network. It is worth noting that Round-Trip-Times (RTTs) from Edge to Core and from Core to Cloud are influenced by the geographical distances, the complexity of traversed networks, and the network congestion that may occur during peak hours. The network architecture shown in Fig. 4 enable the deployment and the placement of the 5G Service Based Architecture (SBA) components needed to lower the latency at the Edge layer. In particular, a User Plane Function (UPF) is required at the Edge layer to forward user-plane traffic to the MEC server as shown in 4. The closer the UPF is to the Next Generation Node B (gNB), the lower the RTT is, so increasing overall system performances. At the same time, placing UPFs at gNB calls to deploy more UPFs so requiring higher control and coordination in the network and may require additional computation and networking capabilities than running one single centralised UPF. Having a number of UPFs also means the system reliability is increased as the loss of one single UPF affects only a small portion of the users. This leads to a trade off between the number of deployed UPF and the computation and network resources to be deployed into the loop. UPFs should be deployed at the edge only when needed to support low latency scenarios. In our reference architecture, we have the UPF at the edge layer, co-located with the Micro Data Center running the MEC Platform. This is a good trade-off as it does not require deploying a number of ad-hoc micro datacenters at the gNB side but it only requires to deploy MEC servers at edge layer of the networks. As long as the gNBs are conencted through low latency and high capacity links (i.e., optical fiber) the system is able to serve low latency

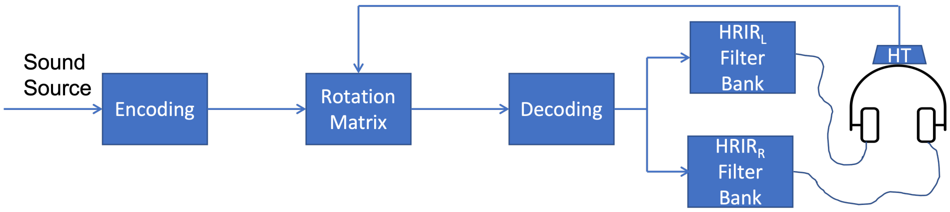


Fig. 3. Virtual loudspeakers based binaural rendering with head tracking.

to the end users. In our application, the traffic from and to the users can follow one of the possible paths depicted in Fig. 4, accordingly to the available network architecture. This depends by the presence of the MEC platform. If a MEC platform is present, it is possible to enable the presented architecture deploying a 5G UPF able to forward the traffic to the proper MEC node.

III. Implementation

Concerning the spatial audio service, we implemented the two solutions described in section II, with Cycling '74 Max software, a graphical software coming from the experimental music world which is now used by many researches in the field of multimedia [20]. It has been created by Miller Puckette who also developed an open source version known as Pure Data [21]. Both of these softwares are written in C.

As previously stated, the most of the computations have to be assigned to the edge server, thus two different patches have been designed, one for the end user device, which is in charge of sending head tracking data, and two patches at the network side, running respectively the classical solution and the virtual loudspeakers based one.

A. Max MSP

1) Client: The client receives the spatialized audio computed as a function of the information related to the position of the head. The client thus sends Head Tracker (HT) data and receives two audio channels. In our implementation shown, the HT acquisition takes place through a Supperware 1 mod head tracker whose parameters are managed by the Bridgehead application [22]. One limitation of Bridgehead is that the maximum data tracking rate is 100Hz. The acquired data is sent from Bridgehead to Max MSP via the UDP protocol to the patch on the server. Fig. 6 shows the client patch and the Bridghead application.

2) Server: Two different patches have been uploaded on the server, corresponding to the two considered methods for achieving binaural rendering. In both cases, the patch on the server takes the information about the current position of the final user's head and uses it either to chose the proper HRIRs for convolution with the sound to be spatialized, or to apply the rotation of the soundfield before Ambisonics decoding. In the case of the classic solution depicted in Fig. 7, Spat5 objects were used. Spat is a real-time spatial audio processor that allows composers, sound artists, performers, and sound engineers to control the localization of sound sources in 3D auditory spaces [23]. External libraries such as HIRT (HISSTools Impulse Response Toolbox) [24], ICST (Institute for Computer Music and Sound Technology) Ambisonic [25] were used in the virtual loudspeakers implementation in Fig. 8. In particular, the [multiconvolve] object capable of convolving at almost 0 latency was used. Comparing the two implementations, one of the most obvious differences is the CPU usage. In the Spat implementation the CPU works much more than in the fixed HRIR implementation.

B. Jacktrip

The audio information is passed from the server to the user via CCRMA's Jacktrip [26]. It is a system for High-Quality, low latency, Audio Network Performance over the Internet via UDP packets. Jacktrip supports bidirectional, uncompressed audio steaming with any number of channels. For jacktrip to work it is necessary that all connected hardware and software must have the same sampling rate (bandwidth) and buffer size (packet size). In our case the sampling frequency was at 44100 Hz and Buffer size at 512 samples. This determines an inherent latency of:

$$\frac{\text{Buffer size}}{\text{Sampling Frequency}} = 11.6\text{ms}$$

It is also important to underline that the quality of the internet service determines audio quality and latency.

C. Network configuration

For this work, we had the chance of exploiting a real 5G network from an Italian operator. The operator provided access at experimental users to a MEC node deployed at 100km distance from the City of L'Aquila. The access to the MEC node was granted to the experimental users utilizing a purposely-built mobile Access Point Name (APN). For the radio segment the experimental users were able to attach to the commercial RAN of the operator, which exploits Dynamic Spectrum Sharing (DSS) on Frequency Division Duplex (FDD) frequency of 1800 MHz.

IV. Results

For evaluating the feasibility of the proposed solution we set up the virtual machine on the MEC server, uploaded the two patches on the server and specified some parameters. In particular concerning the buffer size on QjackCTL it was fixed to 512 samples, because any reduction was causing packet losses. Also the HT rate has been fixed to 100 Hz for latency savings.

Before making any objective evaluation, we decided to proceed with subjective listening tests. They involved two musicians and two engineers as final users, in addition to the authors of this paper. The task was to evaluate on a three level scale (BAD-MEDIUM-GOOD) the perceived quality of the spatialized audio as listened through a pair of semi-open professional studio headphones [27], when the service is implemented at the edge of the network as discussed. The outcoming values have been used as QoE estimation for the service.

When the client is inside the building, many artifacts on the received audio channels could be heard, surely caused by scarce network quality. The overall evaluation on the perceived quality was mostly BAD (6 BAD, 2 MEDIUM). When the client, thus the final user, is moved outside the building, almost in line of sight with respect to the gNB, then the binaural rendering quality becomes extremely appreciable and none was able to listen to any particular artifact on the received sound (overall evaluation was indeed GOOD).

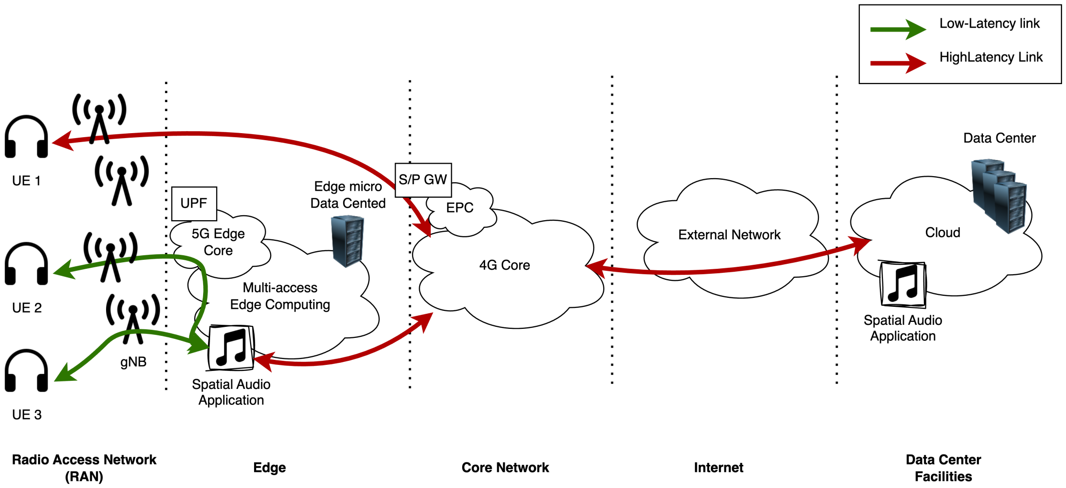


Fig. 4. Considered Multi-access Edge Computing Architecture.

We then tried to evaluate the total Round Trip Time from the HT to the headphones as depicted in Fig. 5.

Since the complete RTT could not be evaluated at the application level due to the temporary unavailabilty of the server for the established testing days, we could give a rough evaluation of the latency by exploiting local data and network latency computations that we preliminarly carried out.

We conducted a preliminary network testing activity to characterize the experienced network performance in common network conditions for the spatial user. We meausured the RTT between a user connected to a commercial MEC-enabled 5G basestation and the closest MEC node. We performed a test in an outdoor scenario with the user connected to a 5G base station at 400m distance. In this scenario we measured an average latency of 29.85 ms with a standard deviation equal to 9.44. We then analyzed an indoor scenario with the user placed inside a building and connected to the same base station approximately at the same distance. In this case the measured average latency was equal to 33.60 ms and the standard deviation 11.56. As expected, the indoor users pays lower performance due to channel penetration impairments. This is confirmed by the measured packet loss that was equal to 2.4% for the indoor user while being 0.9% in the outdoor test.

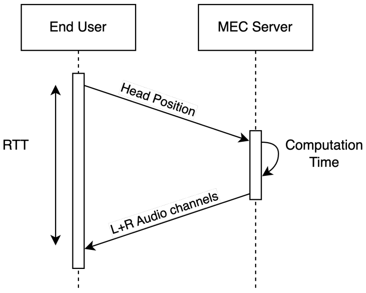
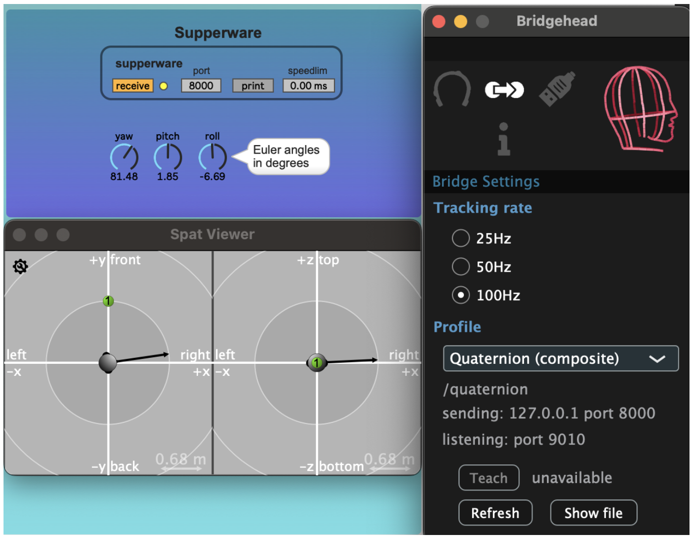


Fig. 5. Sequence Diagram.



8000

0.00 ms

81.48 1.85 -6.69

Bridge Settings

Tracking rate

○ 25Hz

○ 50Hz

● 100Hz

Profile

Quaternion (composite)

/quaternion

sending: 127.0.0.1 port 8000

listening: port 9010

unavailable

Fig. 6. Max MSP and Head Tracking.

Concerning the applications latency we could record the following:

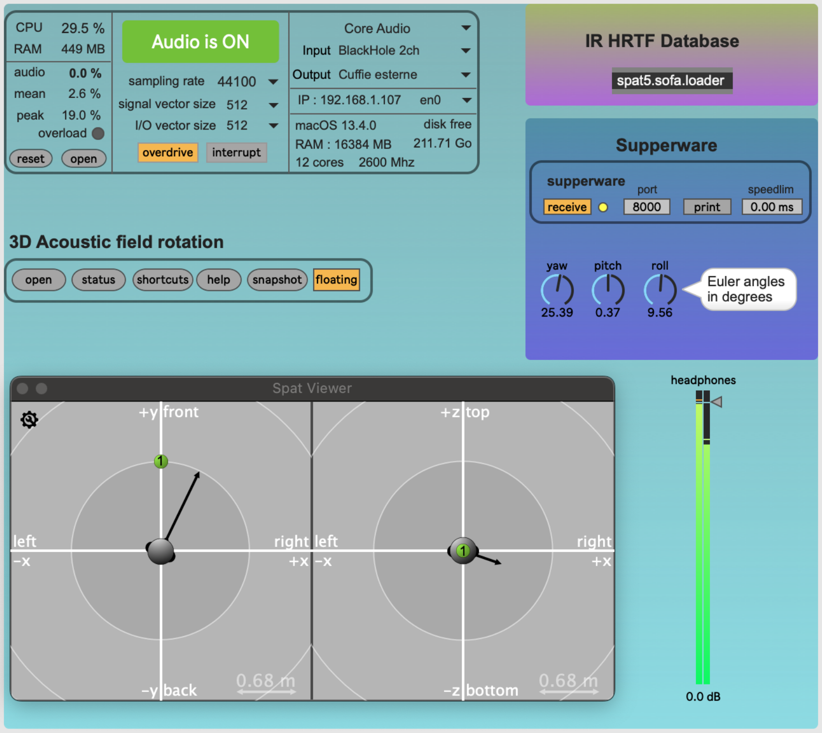
1) HT: $\tau_{HT} = 10$ ms;
2) Binaural Rendering Classical method: $\tau_{sa} \approx 3$ ms;
3) Binaural Rendering Virtual Loudspeakers method: $\tau_{sa} \approx 4$ ms;

The overall latency can thus be approximated as:

$$\tau_{tot} = \tau_{sa} + \tau_{HT} + \tau_{net} \qquad (3)$$

where τ_{net} represents the network latency.

Based on the previously presented data on network latency, and considering that the overall application delay is at most equal to 14 ms, the total latency has a mean of 47 ms indoor and 44 ms outdoor. Both of these latency values are in line with the perceptual tolerability based on [17], but the outdoor



CPU 29.5 %
RAM 449 MB
audio 0.0 %
mean 2.6 %
peak 19.0 %

Audio is ON

sampling rate 44100
signal vector size 512
I/O vector size 512

Core Audio
Input BlackHole 2ch
Output Cuffie esterne
IP : 192.168.1.107 en0
macOS 13.4.0
disk free 211.71 Go
RAM : 16384 MB
12 cores 2600 Mhz

IR HRTF Database
spat5.sofa.loader

Supperware
supperware
port 8000
speedlim 0.00 ms
yaw 25.39
pitch 0.37
roll 9.56
Euler angles in degrees

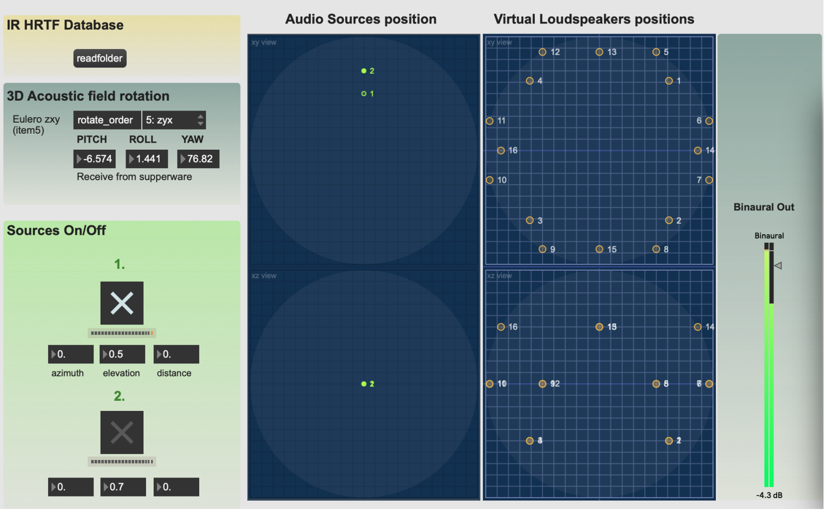
Fig. 7. Max MSP Classical solution spatialization algorithm.

scenario is the only feasible one for the available network due to the lower packet loss and thus lower number of perceived artifacts on the received sound, as previously guessed from qualitative tests.

It has also to be noticed that the two implemented binaural rendering algorithms do not show signficantly important differences in terms of computational delays. We have instead observed a very high CPU burden for the classical implementation, which could not be a problem since we are offloading the computation to the network. What needs to be evaluated is the application latency when more than two sound sources are to be rendered. In this case, we expect that the virtual loudspeakers solution would be more efficient since the number of HRIRs, and thus of convolutions, remains fixed.

V. Conclusions

In this paper we demonstrated the feasibility of exploiting a real MEC configuration to implement a spatial audio service at the edge of the network. Moving from a subjective judgement of the quality of the experience, we computed the overall



IR HRTF Database

3D Acoustic field rotation
Eulero zxy (item5)
rotate_order 5: zyx
PITCH -6.574
ROLL 1.441
YAW 76.82
Receive from supperware

Sources On/Off
1.
azimuth 0.
elevation 0.5
distance 0.
2.
0.
0.7
0.

Fig. 8. Max MSP virtual loudspeakers spatialization algorithm.

latency and demonstrated that under certain conditions it respects the limits of perception tolerability.

Offloading the computation of a spatial audio service at the edge of the 5G network offers several advantages, primarily driven by reduced latency, improved performance, and enhanced scalability. At the same time with edge computing, it is possible to dynamically switch between different spatial audio algorithms based on network conditions, user preferences or the need for synchrony with video, which is in turn particularly important for applications such as augmented reality (AR) and virtual reality (VR). This adaptability ensures that users receive the best possible audio experience, as the system can choose the optimal algorithm for the current scenario in real-time.

We are aware that MEC paradigm is not only moving computation and storage resources to the edge but is also calling for end-to-end network slicing to meet the application level desired Key Performance Indicators. Up to now, telecommunication providers are not able to sell end-to-end network slices up to the final users but there is significant evidence that they will do in the future. We plan to implement end-to-end support within the network owner in the future to demonstrate benefits of this approach in a real network environment.

Future works also include evaluating the potential of the offloading procedure when the number of sources to be rendered increases, in order to give a very first classification of the performance of different spatial audio algorithms on a networked scenario. It is also fundamental to add the simulation of a room with its geometrical and absorbing properties. This is particularly useful in a NMP context to increase immersivity since players may freely decide to share the same (even extreme) acoustic environment, but it introduces important computational issues since in the most basic implementation, each reflection is to be considered as a virtual sound source [28].

For reduction of latency, we are also aware that it is fundamental to remove all the graphics behind Max and maintain only an executable file, as well as working on optimization of parameters such as sampling frequency and Max and Jacktrip vector size. That is why the plan is to arrive to a stable configuration, moving to Pure Data and subsequently work with only the C code.

Another aspect to be considered is the opportunity of allowing the user to chose the final set up of the spatial audio rendering, in such a way that also a specific loudspeakers distribution could properly reproduce the virtual sound sources.

Acknowledgement

The Authors would like to thank Professor Fabio Graziosi for his helpful suggestions and discussions, which have improved and strengthened the presentation of this manuscript.

References

[1] F. Franchi, A. Marotta, C. Rinaldi, F. Graziosi, L. Fratocchi, and M. Parisse, "What can 5g do for public safety? structural health monitoring and earthquake early warning scenarios," *Sensors*, vol. 22, no. 8, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/8/3020

[2] A. Othman and N. A. Nayan, "Public safety mobile broadband system: From shared network to logically dedicated approach leveraging 5g network slicing," *IEEE Systems Journal*, vol. 15, no. 2, pp. 2109–2120, 2020.

[3] L. Turchet and P. Casari, "Latency and reliability analysis of a 5g-enabled internet of musical things system," *IEEE Internet of Things Journal*, pp. 1–1, 2023.

[4] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Toward low-latency and ultra-reliable virtual reality," *IEEE Network*, vol. 32, no. 2, pp. 78–84, 2018.

[5] X. Qiao, P. Ren, G. Nan, L. Liu, S. Dustdar, and J. Chen, "Mobile web augmented reality in 5g and beyond: Challenges, opportunities, and future directions," *China Communications*, vol. 16, no. 9, pp. 141–154, 2019.

[6] J. N. Njoku, C. Ifeanyi Nwakanma, and D.-S. Kim, "The role of 5g wireless communication system in the metaverse," in *2022 27th Asia Pacific Conference on Communications (APCC)*, 2022, pp. 290–294.

[7] D. B. Rawat and H. El Alami, "Metaverse: Requirements, architecture, standards, status, challenges, and perspectives," *IEEE Internet of Things Magazine*, vol. 6, no. 1, pp. 14–18, 2023.

[8] S. Serafin, F. Avanzini, A. De Goetzen, C. Erkut, M. Geronazzo, F. Grani, N. Nilsson, and R. Nordahl, "Reflections from five years of sonic interactions in virtual environments workshops," *Journal of New Music Research*, vol. 49, no. 1, pp. 24–34, 2020.

[9] R. Shukla, R. R. Radu, M. Randler, and R. Stewart, "Real-time binaural rendering with virtual vector base amplitude panning," *Journal of the audio engineering society*, march 2019.

[10] T. Lossius, P. Baltazar, and T. de la Hogue, "Dbap - distance-based amplitude panning," in *International Conference on Mathematics and Computing*, 2009.

[11] D. G. Malham and A. Myatt, "3-d sound spatialization using ambisonic techniques," *Computer Music Journal*, vol. 19, p. 58, 1995.

[12] A. Drioli, "Analysis of Binaural Technology and Surround Rendering for Headphones Reproduction," SAE Institute London, Tech. Rep., 03 2016.

[13] S. Angelucci, F. Franchi, F. Graziosi, and C. Rinaldi, "Binaural spatialization: Comparing head related transfer function models for use in virtual and augmented reality applications," in *Multimedia Technology and Enhanced Learning*, S.-H. Wang and Y.-D. Zhang, Eds. Cham: Springer Nature Switzerland, 2022, pp. 601–613.

[14] L. Turchet, M. Lagrange, C. Rottondi, G. Fazekas, N. Peters, J. Østergaard, F. Font, T. Bäckström, and C. Fischione, "The internet of sounds: Convergent trends, insights, and future directions," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 264–11 292, 2023.

[15] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5g network slicing using sdn and nfv: A survey of taxonomy, architectures and future challenges," *Computer Networks*, vol. 167, p. 106984, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128619304773

[16] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. S. Rummukainen, and E. Habets, "Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information," A. I. C. on Audio for Virtual and A. R. (AVAR), Eds., 2018.

[17] A. Lindau, "The perception of system latency in dynamic binaural synthesis," *Proc. of 35th DAGA*, pp. 1063–1066, 2009.

[18] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, no. 3, pp. 171–218, 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0003682X9290046U

[19] M. Noisternig, A. Sontacchi, T. Musil, and R. Höldrich, "A 3d ambisonic based binaural sound reproduction system," 2003.

[20] C. '74, "What is Max?" https://cycling74.com/products/max, [Online; accessed 19-July-2023].

[21] M. Puckette, "Pure Data (Pd): real-time music and multimedia environment," http://msp.ucsd.edu/software.html, [Online; accessed 19-July-2023].

[22] S. Ltd, "Supperware Head Tracker," https://supperware.co.uk/headtracker-overview, [Online; accessed 20-July-2023].

[23] J.-M. Jot and O. Warusfel, "Spat : A spatial processor for musicians and sound engineers," in *CIARM: International Conference on Acoustics and Musical Research*, 1995. [Online]. Available: https://api.semanticscholar.org/CorpusID:117536243

[24] A. Harker and P. A. Tremblay, "The hisstools impulse response toolbox: Convolution for the masses," in *Proceedings of the international com-*

puter music conference. The International Computer Music Association, 2012, pp. 148–155.

[25] J. C. Schacher, “Seven years of icst ambisonics tools for maxmsp–a brief report,” in *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, vol. 1, 2010.

[26] J.-P. Cáceres and C. Chafe, “Jacktrip: Under the hood of an engine for network audio,” *Journal of New Music Research*, vol. 39, no. 3, pp. 183–187, 2010. [Online]. Available: https://doi.org/10.1080/09298215.2010.481361

[27] “K240 Studio akg professional studio headphones, data sheet,” akg.com/on/demandware.static/-/Sites-masterCatalog_Harman/default/dw4b00c573/pdfs/AKG_K240_Studio_Spec_Sheet.pdf, accessed: 2023-07-12.

[28] 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.