Demonstrating Interoperable Immersive Audio Communication for the Metaverse

Nils Peters[1], Dominik Häußler[2], Clara Romero Soria[3], Matthias Geier[3],
Edwin Mabande[3], Alexander Adami[3], Stefan Döhla[3]
[1]*International Audio Laboratories Erlangen, Friedrich-Alexander-Universität, Erlangen, Germany*
[2]*DSP Solutions GmbH & Co. KG, Regensburg, Germany*
[3]*Fraunhofer IIS, Erlangen, Germany*
nils.peters@audiolabs-erlangen.de

***Abstract*—Effective communication is a cornerstone of many metaverse concepts, underpinning the immersive and interactive experiences they promise. In this demonstration, we will showcase an interoperable low-latency audio communication chain, based on recent 3GPP and MPEG immersive audio standards, in combination with edge-based speech enhancement technologies. We aim to explore advanced immersive communication scenarios among users accessing the metaverse through diverse hardware configurations and device capabilities.**

***Index Terms*—metaverse, communication, interoperability, immersive audio rendering, speech enhancement**

I. Introduction

The metaverse is evolving as a digital platform for social interaction, entertainment, education, and commerce. To create compelling and immersive experiences and to allow users to interact naturally in the metaverse as if they were together in the same physical space, two key elements are high-quality audio communication and the plausible rendering of virtual acoustic environments. However, the current landscape of virtual worlds and various metaverse platforms is fragmented, with limited interoperability. Users are often unable to communicate across platforms which are often tied to specific consumer devices. This breaks the continuity and consistency of the metaverse experience and excludes users with other hardware setups. To realize the metaverse as an open, and interconnected virtual experience, it is critical to develop audio systems that enable interoperable and low-latency audio communication across different hardware setups and that can scale the immersive rendering according to the device's capabilities. Furthermore, to support accessing the metaverse in noisy environments, enhancing the captured speech signals according to the microphone configurations of the input devices is desired. This demonstration shows an interoperable concept for immersive audio communication for the metaverse. Our system leverages recent 3GPP and ISO/IEC MPEG standards and extends them with edge-based speech enhancement technology. While others (e.g., [1]) have proposed immersive audio rendering methods for the metaverse, this demo additionally explores the real-time communication aspects.

The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS.

II. Components

As depicted in Figure 1, the demonstration consists of a two-user scenario. User B is equipped with a 6 DoF head-mounted display (HMD) and a headphone with a built-in microphone. User B can freely navigate with 6 degrees of freedom (6DoF). User A's setup differs and consists of an ordinary 2D video display, stereo loudspeakers, and a far-field microphone array. User A is static and cannot navigate within the scene.

A. Virtual Scene Representation and Immersive Rendering

MPEG-I immersive audio is the emerging ISO/IEC standard for real-time rendering of 6DoF audio for virtual and augmented reality. It supports a wide range of capabilities starting from different audio source types (i.e., channels, objects, scene-based audio (ambisonics), and sources with spatial extent), support of source directivity and Doppler effect, up to more advanced acoustic effects, such as rendering of early reflections, occlusion and diffraction, late reverberation, and many more. The metadata describing the audio scene is efficiently coded into a bitstream.

The scene context for this demo comprises a restaurant. Here, we envision that User A represents the barkeeper behind the counter and User B is a guest who can freely navigate within the scene. Figure 2 shows the visual rendering of the scene within Unity which is used to represent the visual components via a traditional TV flat-screen to User A and via HMD to User B.

The audio scene is encoded as an MPEG-I immersive audio scene [2]. It consists of various ambient sound sources one would expect from a restaurant as well as a model of all acoustically meaningful surface areas and their acoustic properties, (e.g. absorption, transmission, etc.) to auralize acoustic effects, such as early reflections, diffraction, and occlusion. Additionally, when User B communicates with User A, User B will hear his own voice rendered within the virtual scene. This is possible since MPEG-I allows feeding locally captured audio into the renderer and reproducing it with the corresponding acoustic effects of the scene.

B. Voice Communication

This demo uses the Immersive Voice and Audio Services (IVAS) codec, recently standardized by 3GPP [3], for low-

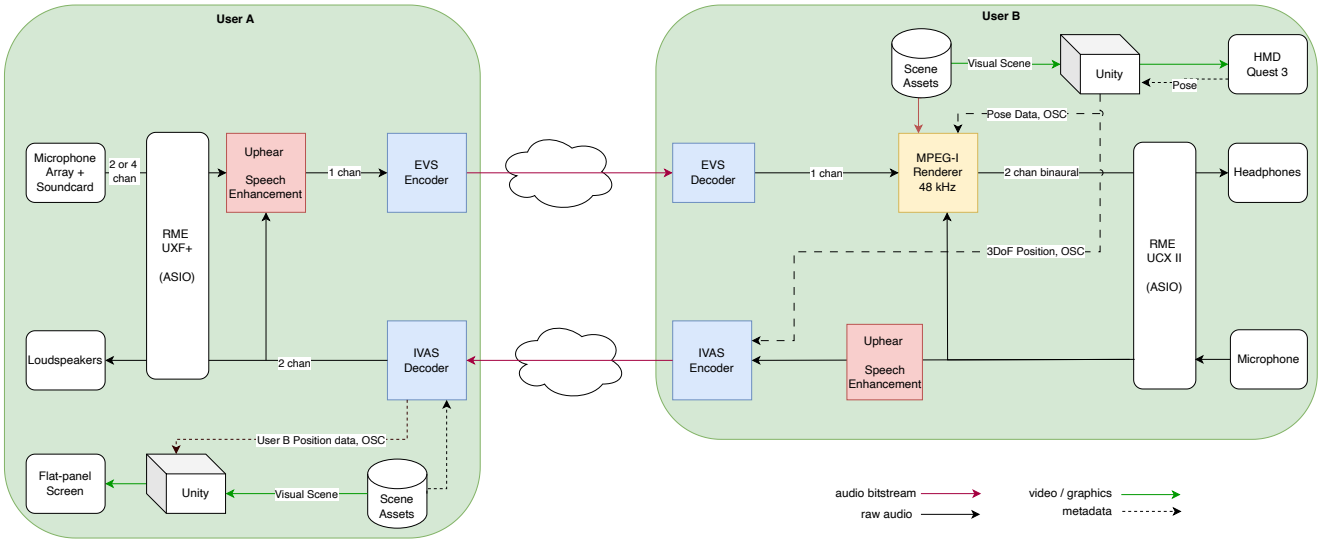

Fig. 1. Hardware components and signal flow of the demo

Fig. 2. Unity visualization of the demo scene

latency audio communication between the two endpoints. It supports the coding of immersive audio in various formats, including stereo, scene-based audio (Ambisonics), and object-based audio; it is based on the 3GPP EVS codec [4]. Within this demo User A sends a mono speech signal to User B using the EVS mode inside IVAS which is then rendered by the MPEG-I renderer together with User B's voice. In the opposite direction, object-based audio, i.e. mono audio plus metadata describing the spatial properties, is encoded, transmitted, decoded, and rendered for User A while also taking the MPEG-I scene into account. For this, the room reverb is converted from MPEG-I representation to the IVAS representation to avoid another rendering stage.

C. Speech Enhancement

Besides low latency, speech intelligibility is another important requirement for effective audio communication. We improve speech intelligibility via the edge-based Uphear Voice Quality Enhancement (VQE) technology before the transmission. Uphear VQE is an advanced acoustic front-end that includes classical and AI-based features such as echo control, noise reduction, dereverberation, automatic gain control, and target speaker extraction.

III. Conclusion

We demonstrated the potential of advanced audio communication technologies for the metaverse. By leveraging recent 3GPP and MPEG immersive audio standards, combined with edge-based speech enhancement, we have illustrated an interoperable low-latency audio communication and immersive rendering chain. This system strives to be adaptable across various hardware configurations and device capabilities, addressing the diverse needs of metaverse users. In the future, we plan to extend the demo to evaluate other technologies regarding their use within the metaverse.

References

[1] J.-M. Jot, R. Audfray, M. Hertensteiner, and B. Schmidt, "Rendering spatial sound for interoperable experiences in the audio metaverse," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2021, pp. 1–15.

[2] J. Herre, S. Disch, C. Borss, A. Silzle, A. Adami, and N. Peters, "MPEG-I immersive audio: A versatile and efficient representation of VR/AR audio beyond point source rendering," in *Proc. of the AES 6th International Conference on Audio for Games*, Tokyo, Japan, 2024.

[3] "3GPP TS 26.250 18.0.0: Codec for Immersive Voice and Audio Services (IVAS); General overview," 3rd Generation Partnership Project (3GPP), Technical Specification TS 26.250, 2024. [Online]. Available: https://www.3gpp.org/DynaReport/26250.htm

[4] "3GPP TS 26.441 18.0.0: Codec for Enhanced Voice Services (EVS); General overview," 3rd Generation Partnership Project (3GPP), Technical Specification TS 26.441, 2024. [Online]. Available: https://www.3gpp.org/DynaReport/26441.htm