Human Interaction and Emerging Technologies (IHIET-AI), Vol. 201, 2026, 14–21
https://doi.org/10.54941/ahfe1007156

On Immersivity of Transmitted Spatial Sound for Human-Machine Interaction

Jan Holub and Jakub Turinský

Department of Measurement FEE CTU, Prague, Czech Republic

ABSTRACT

The paper presents the results of extensive subjective tests that clarify the relationship between head-tracking delay and the quality of subjective spatial immersive experience in various tested situations. The results can be directly applied to specify the minimum technical requirements for designing AI-based audio-immersive communication and control systems.

Keywords: Audio man-machine communication, Subjective testing, Head-tracking, Immersiveness, Quality of experience

INTRODUCTION

Spatial sound perception is a natural way of perceiving sound by humans, serving for spatial orientation, escape from imminent danger, or shifting attention toward an object of interest. In the modern world, spatial sound, transmitted, for example, via telecommunications networks, is increasingly used to convey an additional dimension of information to the recipient (e.g., an air traffic controller can virtually hear a pilot from the direction where their aircraft is located, etc.). Given the expanding field of AI for generating spatial information, it is necessary to investigate the influence of basic technical parameters on the subjective experience of the recipient, represented, for example, by the immersiveness of communication, the intelligibility of transmitted speech, spatial mix, or the spatial resolution of multiple sound stimuli. When using AI to generate spatial sound, the influence of (computational) delay on the quality of the subjective experience is crucial (the recipient's head movement is tracked using head-tracker embedded in the headphones, and the audio channels information is rendered in real time to compensate for any head movements of the subject, so that the sound appears to come from a fixed location regardless of the head's orientation). Any delay in this computational loop can compromise the subjective perception or quality of communication.

STATE OF THE ART

Immersion plays a crucial role in enhancing advanced audio-visual media experiences, enveloping consumers in a captivating scene. Immersive audio, characterized by a sense of 'being there,' is already prevalent in various domains such as cinema, home theatres, music streaming, virtual reality, and

Received August 15, 2025; Revised November 28, 2025; Accepted December 3, 2025; Available online April 1, 2026

gaming. In contrast, mobile communication methods like telephony and conferencing primarily rely on monaural audio experiences, lacking any spatial dimension. This remains true despite advancements such as the AMR-WB codec for HD Voice and the EVS codec for HD Voice+. Bridging this gap in immersive audio could rejuvenate mobile communications. When paired with appropriate video elements, immersive mobile communications will pave the way for innovative services and distinct telepresence-like experiences (Bruhn, 2025).

Immersive audio can be represented in several formats, including channel-based, scene-based (Ambisonics), and object-based audio. In a more defined context, an immersive format typically requires head-trackable 3D audio, allowing listeners to experience a fixed audio scene while changing their head orientation. Even in cases where head-trackability is not essential, stereo audio (also referred to as binaural audio) can also be viewed as immersive. Numerous audio codecs support these immersive formats, particularly for streaming, broadcasting, and storage. Notable examples include MPEG-H 3D Audio, ETSI AC4/Dolby Atmos, and DTS UHD. For conversational services, codecs such as ITU-T G.711 and G.722D stereo extensions, along with IETF Opus—which also supports multichannel and Ambisonics—are available. However, none of these codecs are specifically designed for immersive mobile communication over 4G and 5G cellular networks. There is a need for a codec that meets stringent requirements for latency, compression efficiency, and resilience to packet loss, alongside audio front-end and back-end functionalities that can navigate dynamic and unpredictable environments such as homes, offices, conference rooms, and vehicles.

The new IVAS standard (ETSI TS 126 250, 2024) from the 3rd Generation Partnership Project (3GPP) uniquely addresses this requirement by focusing on immersive audio for 3GPP conversational applications. In addition to supporting common immersive audio formats, it introduces a novel audio hybrid waveform and parametric audio representation tailored for mobile phone capture, called Metadata-Assisted Spatial Audio (MASA). Furthermore, IVAS provides a comprehensive framework for mobile communication, which includes Discontinuous Transmission (DTX) with Voice/Signal Activity Detection (VAD/SAD), Comfort Noise Generation (CNG), packet loss concealment, integrated Jitter Buffer Management (JBM), and an integrated renderer with default binaural filter sets, as well as split rendering capabilities for head-tracked binaural audio on lightweight devices.

Standardized as part of Release 18, the IVAS codec serves as a foundation for not only conversational voice services but also for innovative extended reality (XR) communications, as well as live or prerecorded streaming and messaging services. Use cases encompass sharing immersive experiences in telephony and conferencing, capturing and transmitting immersive scenes in ad-hoc settings, enhancing in-car communication and sound experiences, as well as XR applications such as virtual conferencing, streaming, and content distribution.

MOTIVATION

As using a head-tracker embedded in headphones principally requires backward communication between the device and the playout platform (e.g., smartphone), where the split-rendering calculation will have to be performed, typically with strong constraints on processing power and battery consumption, a non-negligible delay between the listener's head movement and proper, spatially repositioned acoustic information delivery to the listener's ear (see Fig. 1) can be expected (Rämö, 2025).

Any noticeable delay can severely impact the subjective immersiveness, negatively affecting the user experience (UE). At the same time, any effort to minimize this delay will require an increase in computational resource demands or a different wireless protocol for use between the playout platform and the headphones with a head tracker, which would increase equipment costs and power consumption.

Therefore, the dependency of subjective perception of delay between head movement and proper acoustic rendering is of great interest (Meyer-Kahlen et al., 2023). Obviously, this subjective influence will heavily depend on the application or use case where the head tracker is used - compare, for example, listening to background music while working on a different matter versus listening to immersive spatial recording of orchestra music with full concentration on the music (Stitt et al., 2016). Another application area where minimum delay will be required is any potential professional use of the IVAS (e.g., pilots, airport approach control dispatchers, etc.).

METHODOLOGY

Test Setup

The test was designed as four identical test setups, each at a separate table, inside a laboratory. Each station had prepared different tracks with different prerecorded French speech monologues in Reaper digital audio workspace and consisted of hardware: a personal laptop, soundcard Audent EVO 4, Supperware Headtracker 1, and Shure SRH840A headphones. For each track consisting of different prerecorded French speech monologues, there was a predefined delay in milliseconds in a built-in delay plug-in.

The test consisted of a reference audio track with minimal latency and eight tracks for the user to evaluate, one of which was also with minimal delay to check if subjects could spot the unaffected audio. The subjects listened to the audio tracks one after another, switched between them manually, and filled in the evaluation results on a paper form provided to them after each track. The set of tracks and the sequence of the delays were identical for each test station. The order delay sequence was selected with logarithmically increasing steps, ranging from 0 to 500ms, and in a random order to ensure an unbiased evaluation. More on the testing process in 4.4 Test flow. The minimal latency (reference track) was 3.4ms, which was the minimum that the SW and sound card combined could deliver.

The participants were instructed to perform a specific set of movements during the listening phase (see Chapter 4.4, Test flow).

The volume of the audio was set to a common level (ITU-T P.800, 1996) of 73 dB SPL(A), corresponding to –26 dBoV in the digital recording.

Test Environment

The tests were held in an office-like laboratory environment. In the room, there were four individual tables arranged in a circular formation, at least 1.5m apart, so that the participants could not see or hear each other, thereby avoiding potential influences. On each table, the test equipment was positioned, creating four test stations for the participants.

The listening and headtracking equipment consisted of closed headphones (Shure SRH840A-EFS), a sound card (Audent EVO 4), and a Supperware Headtracker 1, positioned on the headphones' headband. The testing equipment also carried a laptop on which the test was prepared as a session in a DAW program. After being instructed, participants played each track individually, which had its own hidden delay implemented.

To collect participants' responses, printed questionnaires were used to avoid confusion with electronic devices.

Test Subjects

Ten native French-speaking participants were randomly selected, between the ages of 18 and 30 years old, including both males and females. The subjects were asked to self-report about their hearing ability and other relevant details.

Subjects were introduced to the theory of binaural audio with headtracking before the session started, unaware of the exact parameters (latency) we were going to test, and asked for their general opinion of the listening experience.

Test Flow

The overall test setup proceeded as follows: up to four participants were introduced to the spatial audio with head-tracking technology and positioned in front of their identical test setup. The head tracker was calibrated to their approximate fixed position on the chair, with their head position facing the laptop. Then, they were presented with a test consisting of two phases, labelled as Immersive scenario 1 and Immersive scenario 2, each followed by a paper form with a short introduction and instructions for the head movements. They were asked to rotate their head during the sample listening: first, right, wait for four seconds, then rotate left, wait for another four seconds, and return to the neutral, forward-facing position, holding for four seconds. After that, they were able to move freely until the sample was finished. The paper for each phase of the test also consisted of a table with a set of four questions for each track, for them to evaluate by grades 1, 2, 3, 4, and 5 (with 1 being the worst and 5 being the best grade).

In the first phase, they were asked to evaluate the spatial audio experience in general, without mentioning the parameters affecting each track. They were asked to answer four different questions:

- Could you easily locate the sound source in space?
- Did the sound feel natural in relation to your movements and position?
- Were there any moments when you felt the sound source change position without reason?
- What was the overall quality of the sample and your immersive experience?

In the second phase, the process repeated itself identically, with the only change in the questions now consisting of direct questions about latency and its impact on their immersive listening experience:

- Did you perceive any delay between your head movement and the "movement" of the sound?
- If there was a delay, did it seem disturbing to you?
- Does perceived latency affect your immersion or listening comfort?
- How did the delay affect the overall quality of the sample and your immersive experience?

Similar to ITU-T P.835, the first questions were used to focus subjects' attention on the tested aspects of the audio stimuli, and only answers to the last (general) question were evaluated statistically, as described further.

Each phase contained a set of 9 samples. First, the sample labelled as Sample 0 served as the reference sample for the user to gain an understanding of how the system works when the latency is ideal. The rest consisted of a set of 8 different delays randomly chosen from the selected sequences of delays. One of the eight samples was again a reference track with minimal delay to see if the user could spot it and ensure the subjects were actually evaluating the error affecting the samples. The pseudo-randomized sequence presented to the subjects follows:

- General question (phase one - Immersive scenario 1) delay sequence: 85, 45, 130, 3, 253, 503, 170, 380 ms.
- Delay question (phase two - Immersive scenario 2) delay sequence: 130, 253, 3, 85, 503, 170, 380, 45 ms.

Each participant began with the first phase, played each track once, and moved accordingly during the listening. They then graded each sample using the mentioned four questions. Then they moved to the second phase, which went identically. The entire test took under 20 minutes to ensure the participants could maintain their focus and concentration.

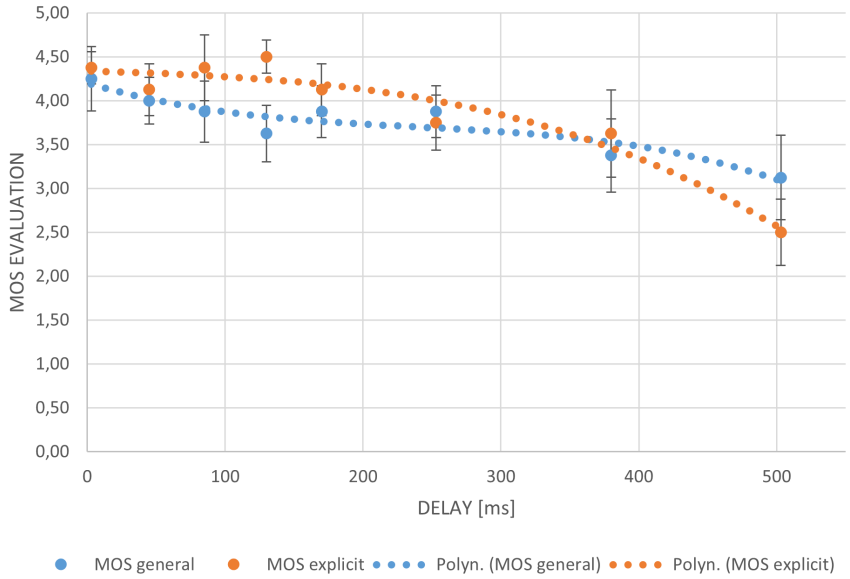

Figure 1: Subjective test results – MOS for general quality and delay-explicit assessment.

RESULTS

The participants' grades, on a scale of 1–5 (with 1 being the worst and 5 being the best grade), were averaged for each sample (specific delay) to obtain the Mean Opinion Score (MOS). The averaging was done across all ten participants for each phase of the testing. The processed questions were the final questions of each phase. For the first general phase, "What was the overall quality of the sample and your immersive experience"? While for the second phase (where participants knew the tested impairment), "How did the delay affect the overall quality of the sample and your immersive experience"? Overall, ten participants were tested, of whom two were discarded since their evaluation did not change with the change in delay, indicating they had not understood the task or recognized the impairment properly.

Table 1: Subjective test results – MOS for general quality and delay-explicit assessment, including statistical evaluation.

Delay [ms]	MOS General	MOS Explicit	RMSE (MOS General)	RMSE (MOS Explicit)	T-test
3	4,25	4,38	0,37	0,18	0,208079
45	4,00	4,13	0,27	0,30	0,201983
85	3,88	4,38	0,35	0,38	0,014138
130	3,63	4,50	0,32	0,19	0,000152
170	3,88	4,13	0,30	0,30	0,066985
253	3,88	3,75	0,30	0,31	0,219247
380	3,38	3,63	0,42	0,50	0,156771
503	3,13	2,50	0,48	0,38	0,011569

The results are summarized in Table 1 and Figure 1, where the MOS scores for the different tested delay values are shown, along with their corresponding standard deviations (represented by error bars in Figure 1).

DISCUSSION

The dependency of the quality on tracker delay for the generic question (without mention of delay) and for the question with explicit mention of delay as an influencing factor has a similar overall trend, with a downward tendency for higher delay values. However, they differ statistically significantly ($p < 0.05$) in two ranges of delay:

For values of 80 and 130 ms, the MOS scores for the explicit question about the influence of delay on overall quality are significantly higher than those for the generic question. It is likely that explicitly drawing the test subject's attention to the aspect of delay as a tested factor distracts the subject, and the assessment of overall quality is then higher (less critical).

Conversely, for the highest tested delay value (500 ms), the quality rating using the generic question (without explicitly mentioning the delay) is statistically significantly higher. For such a high delay value, drawing attention to the evaluated parameter is apparently associated with an awareness of the very low practical usability of such a transmission.

In follow-up experiments, it would be appropriate to perform measurements for higher delay values, e.g., in the range of 0.5–1s, and also to reduce the variance of the MOS rating by using more test subjects.

CONCLUSION

The experiment compared two methods of subjective testing of the delay between transmitted sound and compensation for the user's head movement using a head tracker and a feedback renderer.

The measured results guide developers of spatial encoders, head trackers, and other elements of the transmission chain, e.g., for communication (human-machine), for the appropriate choice of testing methodology for subjective verification of the resulting design or its optimization.

ACKNOWLEDGMENT

The authors would like to thank Buvat de Virginy Arthur and Benseghir Younès for preparing the questionnaires, gathering the participants and managing the test flow.

REFERENCES

Bruhn, Stefan (2025) "3GPP IVAS Codec – Perspectives on Development, Testing and Standardization," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10889939.

ETSI TS 126 250 (2024), Codec for Immersive Voice and Audio Services, General overview, Sophia Antipolis, France.

ITU-T P.800 (1996) "Methods for subjective determination of transmission quality", ITU-T Geneva 1996.

Meyer-Kahlen, Kastemaa, Schlecht, Lokki (2023) Measuring Motion-to-Sound Latency in Virtual Acoustic Rendering Systems. Journal of the Audio Engineering Society. 71. 390–398. 10.17743/jaes.2022.0089.

Rämö, Toukomaa (2025) "Subjective Voice Quality of the IVAS Codec," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10890848.

Stitt, Messonnier, Katz (2016) The Influence of Head Tracking Latency on Binaural Rendering in Simple and Complex Sound Scenes. Audio Engineering Society, New York.