Immersive Voice

Entering the era of Spatial Audio Communication – Implications and opportunities for industry and consumers

White paper

© 2024 Nokia

Contents

© 2024 Nokia

1 Introduction

Voice is our primary means of communication, and telephony has enabled us to connect with voice for over a century. The phone call as we know it has evolved from analog to digital, from fixed to mobile, and from low speech quality to natural speech quality. In recent years we've seen a major transformation in the use of the voice modality – from voice-driven user interfaces to generative AI and extending a traditional phone call or conference call to using voice in chats, games and stories. One major advancement, however, was still lacking: how to enable a fully authentic, immersive sound to be transmitted in these use cases.

The term *spatial audio* may bring to mind movie theatres and soundtracks, where sound sources smoothly transition from left to right, front and back, and even above us. Or video gaming with headphones that can replicate explosion and footstep sounds so realistically that players can become fully immersed in the game world.

In everyday situations, people can perceive sounds coming from different directions. This *outside the head* listening experience is called spatial audio. Simply put, everything that we hear around us is spatial audio.

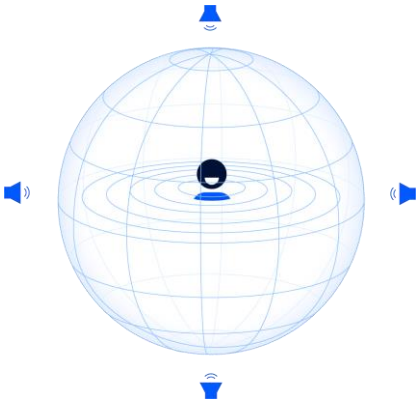


Figure 1: Spatial audio creates a 360° audio experience for the listener.

Experiencing spatial audio gets a bit more complicated when you put on headphones. Music, video soundtracks, and voice calls are typically monophonic or stereo audio and they are experienced *inside the head*, somewhere between the left and right ears.

When wearing headphones your brain needs to be convinced that the sounds you hear are not coming directly from the left or right loudspeaker of your headphones but are instead spatially positioned around you. The illusion of spatial sound is reproduced by modeling human spatial hearing and applying digital filters, so called Head-Related Transfer Functions (HRTFs).

Sound designers and audio engineers in game companies, music or movie productions have the capability and equipment to craft stunning spatial listening experiences. But is it possible for everyone to create spatial audio with the devices and software we already own and use daily? And why would people want to do so?

© 2024 Nokia

To answer these questions, we need to consider the major shifts in how people express themselves using real-time spatial multimedia to aid natural communication between two or more participants. We refer to this as 'Spatial Audio Communication' – where spatial audio becomes an integral part of how we communicate, beyond just entertainment.

This white paper will discuss these industry-redefining topics and explain what Nokia's role is in enabling immersion when watching and listening to smartphone-recorded or live-streamed videos or making a call or messaging with the new voice technology. We will delve into the fundamentals of capturing, transporting, and experiencing spatial audio in real time and offline. Most importantly, we will explore how spatial audio will change the way we communicate in cellular voice calls, in over-the-top services and eXtended Reality (XR).

© 2024 Nokia

2 Spatial audio capture by mobile devices

Spatial audio can be created in professional or home studio software using mixing tools that direct sounds into a desired multi-loudspeaker channel. There are also formats that enable sound “placement” at precise coordinates and moving them around the listener. Playback equipment then renders these sounds based on the number and location of loudspeakers in the space.

Game engines use somewhat similar approach for placing sounds around the listener during gameplay. However, these engines must also be able to render the entire audio scene in real time as players often enjoy free movement within the game world (6 degrees of freedom).

A common aspect of sound design in both movies and games is that soundtracks and audio clips are pre-recorded, even if they would be mixed and rendered interactively like often done in modern games.

However, creating spatial audio doesn't necessarily require dedicated studio equipment.

In fact, to capture spatial audio you only need a device with two or more integrated microphones and a suitable audio recording software. Indeed, just as two ears are sufficient to experience spatial audio, two microphones can adequately capture it. However, more microphones can enable more capabilities.

Nokia has a longstanding legacy in developing spatial audio technologies for mobile devices. OZO Audio software, for instance, has enabled the recording of spatial audio across devices of varying shapes and sizes. This advanced *parametric audio* processing technology is the most suitable approach for capturing spatial audio in mobile devices.

One of OZO Audio’s biggest advantages is that the spatial audio output can be stored into a normal stereo channel, making it accessible on any audio player. Additionally, *spatial metadata* can be generated and stored alongside the audio. This metadata describes, for example, the direction of the sounds and can be used to enable head-tracking of the audio during playback. OZO Audio also provides functionalities that go beyond what is typically available with mono or stereo recording solutions, such as audio zooming and tracking of sound sources. OZO Audio is used in various device types in the market, for example smartphones and tablets, VR cameras, wearable cameras, and consumer cameras.

2.1 Nokia Immersive Voice

Whereas OZO Audio is capable of capturing spatial audio for recording and streaming use cases, Nokia has used it as a basis for an end-to-end spatial audio call technology, *Immersive Voice*.

© 2024 Nokia

Nokia Immersive Voice enables various new conversational use cases for mobile devices. The captured spatial audio is encoded with selected audio codec, transmitted to the other caller(s), and on the receiving side the audio is rendered for playback. For each user it is possible to hear sounds from the other side realistically, being surrounded by the audio scene in real time. The solution is equipped with a set of dynamic controls that can be used to adjust the amount of captured ambience (which sometimes is just background noise), choose the orientation of the audio scene (e.g., selfie/main camera direction), and to enable head-tracking, to name few.

At the heart of the Immersive Voice there are two software development kits (SDKs):

Immersive Voice Client SDK

This software library can be integrated into multi-microphone devices, or into applications that are running on compatible devices. Once accessing microphone signals, it provides live spatial audio for communication apps. It also enables immersive playback experiences, e.g. supporting head-tracking for headphone playback, and stereo widening for device's stereo loudspeaker playback.

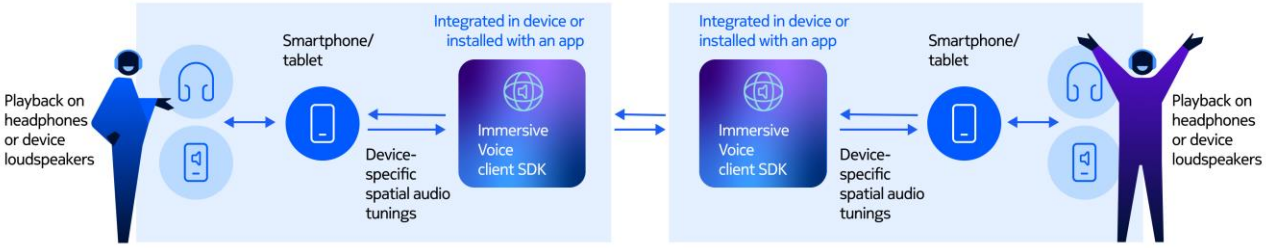


Figure 2: Immersive Voice end-to-end experience.

Immersive Voice Mixer SDK

Integrated into communication services, this software library spatially mixes audio from multiple sources for each user and enables group calls. Users can perceive sounds from various directions and interact dynamically, such as repositioning sound sources within their environment.

© 2024 Nokia

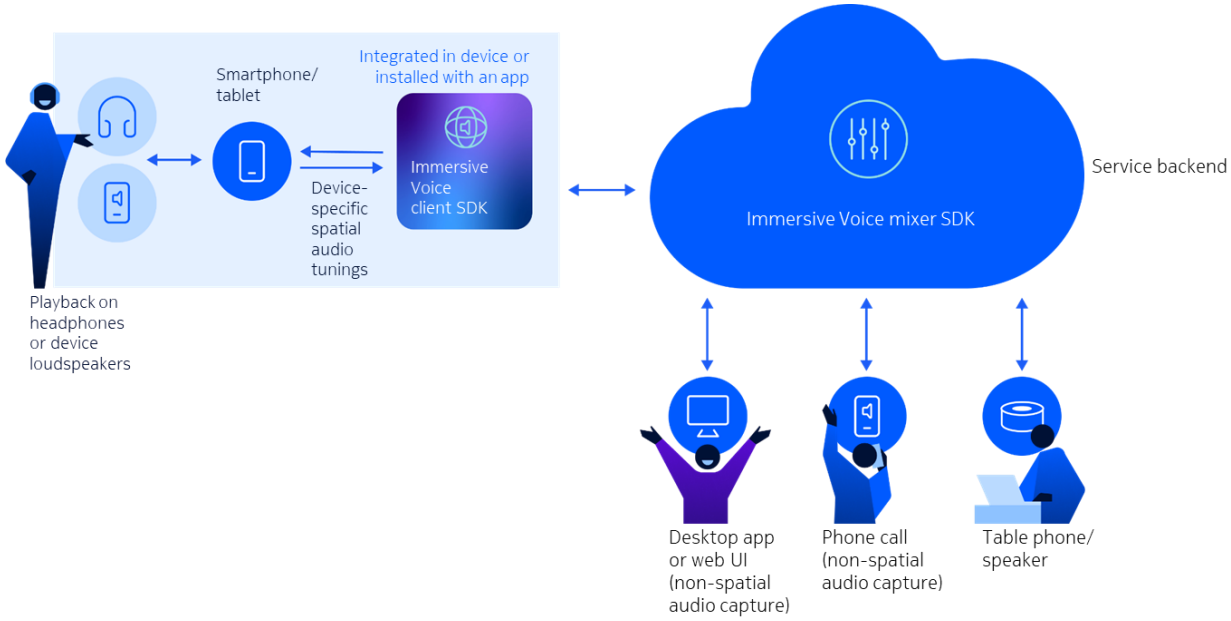


Figure 3: Immersive Voice group call experiences.

2.2 Developing applications with Nokia Immersive Voice

Immersive Voice provides an all-in-one solution for recording voice and video messages with spatial audio, sharing immersive live streams, and participating in 1-to-1 and group calls.

The Immersive Voice Client SDK can be integrated into smartphones, tablets, vehicles, headsets and other types of terminals to enable live capture and playback of spatial audio experiences. It can also be integrated into software applications that are running on the aforementioned devices. Nokia's Immersive Voice Mixer SDK enables spatial group calls that can be scaled from industrial private networks to the consumer use cases of mobile operators and social media services.

Whereas the recorded spatial audio can be stored and shared in formats like AAC, the real-time streamed audio and Voice-over-IP calls require a codec that can support good quality audio over connections of varying bitrates. Opus is one such codec, that is widely supported in the market. Opus can also be supported by the Nokia's Immersive Voice solution.

2.3 3GPP IVAS

To enable spatial audio calls becoming supported also in mobile phone calls, and not only in Over-the-Top (OTT) services, there has been a need to agree on a new voice codec standard in the 3rd Generation Partnership Project, 3GPP.

© 2024 Nokia

Nokia has been a leading contributor of the proven mobile-centric audio coding principles of the new 3GPP standard, the *Immersive Voice and Audio Services (IVAS)* codec. The codec was included in the 3GPP Release 18. As an extension of the widely deployed monophonic Enhanced Voice Services (EVS) codec, IVAS provides full backwards compatibility ensuring interoperability with existing voice services.

A new parametric audio format was developed during IVAS standardization that is particularly suitable for challenging device form factors such as mobile devices. This format is called Metadata-Assisted Spatial Audio, MASA. IVAS codec has an integrated renderer for head-tracked binaural and multi-loudspeaker playback from the MASA format.

The Nokia Immersive Voice Client SDK can be used as an IVAS front-end to capture spatial audio from device microphones into MASA format.

Further information on IVAS can be found, e.g., in this article:
https://www.3gpp.org/technologies/ivas-highlights

© 2024 Nokia

3 Experiencing Immersive Voice through headphones and loudspeakers

Immersive Voice for headphones

Headphones are the optimal way of experiencing spatial audio, as through them it is possible to precisely manage some of the complexities of human hearing. This is achieved through a processing known as binauralization. It essentially creates an auditory illusion by simulating a three-dimensional perception of sound.

Head-Related Transfer Functions (HRTFs) are mathematical filters that play a pivotal role in binauralization. By taking into account the effect of the shape of our head, ears and torso, the filtering is essential for our ability to perceive and localize sounds in a three-dimensional virtual space. By analyzing the subtle differences in sound arrival time and intensity between the ears, our brain creates a sense of auditory spatial awareness. HRTFs are widely used in audio engineering, virtual reality, and spatial audio technologies to create more immersive and realistic sound experiences.

The Nokia audio team have carried out extensive measurements and modeling of HRTFs and we deploy optimal HRTF filters in our binauralization algorithms. With Immersive Voice calls, binauralization is applied in real-time, which is crucial for preserving immersion, especially when applying head-tracking.

Immersive Voice for mobile device loudspeakers

A mobile device needs at least two integrated loudspeakers to produce immersive playback. These twin loudspeakers facilitate direct stereo content playback and, with appropriate rendering processing, can provide an immersive experience around the listener.

The mobile form factor imposes constraints on the size and location of microphones and loudspeakers. The environment of use cases is often dynamic – people move around and the device is used in different orientations – so flexibility is crucial. Therefore, the audio capturing and playback capabilities must be tuned for each device and their use cases.

© 2024 Nokia

Unlike headphones, where each ear receives a dedicated processed signal, all the loudspeakers can be heard by both ears. Despite this, Immersive Voice can still provide an immersive listening experience using an advanced technique called stereo widening.

Especially with symmetric loudspeaker setups, Immersive Voice processes a stunningly wide stereo image extending the physical loudspeaker locations, approaching the level of immersion offered by headphone playback.

In addition to headphone or stereo loudspeaker setups, immersion and Spatial Audio Communication can be created and enjoyed in a wide variety of devices and usage environments with different types of audio capture and playback systems. Besides mobile devices, the environments where we envisage Spatial Audio Communication making a significant impact include cars, offices, living room media systems, meeting rooms and mission-critical systems.

© 2024 Nokia

4 Overcoming the challenges of Spatial Audio Communication

There were a number of challenges to be considered and resolved in order for Immersive Voice to become a robust spatial audio solution.

The term Spatial Audio Communication refers to the utilization of spatial audio - as well as other technologies such as video - in exchanges between people, both in real-time and non-real time. For example, instant voice messaging and stories have become common ways for people to express themselves. This is why one of the key targets in our research and development of new audio solutions has been to achieve more natural communication and interaction for both real-time and non-real time use, whether it is audio-only or includes other modalities such as video or haptics.

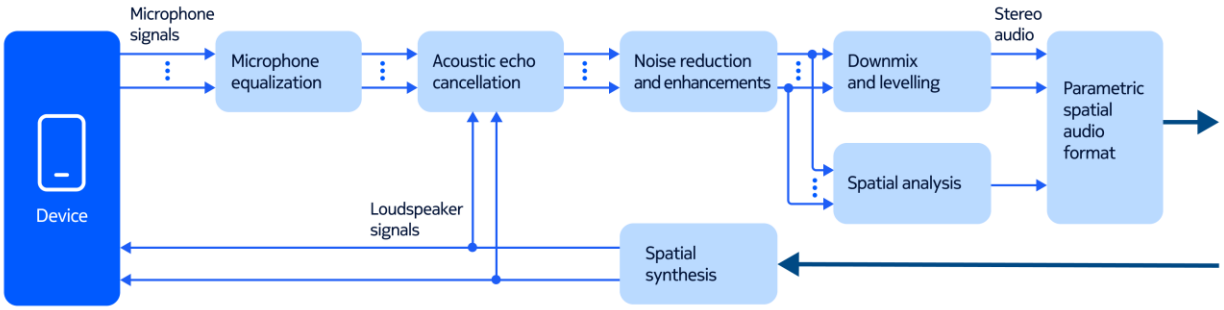


Figure 4: Spatial audio processing in real-time communication.

Spatial analysis and synthesis

Parametric spatial audio capture techniques can be utilized for a wide range of device form factors with different microphone number and placement.

Immersive Voice utilizes algorithms that analyze spatial information from the captured microphone signals based on knowledge of the device size and shape, and the placement and other properties of the microphones themselves. A varying number of microphones is supported, and the algorithms can be easily adapted to different device configurations.

The number and placement of microphones in the capture device define which orientations and dimensions can be spatially analyzed. For example, properly placed three microphones in a smartphone can be used for creating horizontal level (360 degrees in 2D plane) spatial audio, whereas four microphones can create full 3D space and also support portrait orientation.

Unlike traditional multi-channel-based configurations, parametric audio allows for more adaptable rendering of spatial audio for different types of devices. Nokia Immersive Voice

© 2024 Nokia

technology only needs two channels of audio plus spatial metadata to decode high-quality spatial audio.

Depending on the configuration, Immersive Voice can synthesize the captured spatial audio into a binauralized representation already before passing it forward. This can be especially useful if the receiving side does not have a need or capabilities for head-tracking, or the transmission channel does not support the passing of synchronized metadata.

Spatial audio may be synthesized also from one or more mono (or "object") sound sources. The Immersive Voice mixer is able to provide an individual spatial mix of object sound sources for each listener. It can also mix together object sounds and spatially captured sounds and enable real-time interactions, such as moving the sound sources around and support for head-tracking for everyone individually.

Acoustic echo cancellation

When the playback is active, the number and placement of the loudspeakers also affect the spatial audio capture. The acoustic echo caused by microphones capturing the playback signal is a major challenge in all calls, but especially so with Immersive Voice when multiple microphones and multiple loudspeakers are used. Solving this problem has required the development of a novel machine learning-based Acoustic Echo Cancellation (AEC) solution for Immersive Voice.

Real-time conversational audio poses challenges for mobile communication and AEC, especially when the playback happens through the device loudspeakers. In recording and streaming use cases the echo is not such a problem as the audio flow is one-directional.

The target of AEC processing is to remove the loudspeaker signal component from the microphone signals based on a reference signal. In the case of stereo playback both stereo channels are needed as reference input for the AEC.

A traditional solution has been to use a linear AEC filter, which is followed by residual echo suppression. In Immersive Voice, machine learning techniques have been used to create an AEC processor that works for both mono and spatial audio signals.

Noise Reduction

Some use cases – such as a concert or a nature setting – may demand the full sonic ambience. But when the focus is on speech, it becomes necessary to suppress background noise.

Background noise reduction can be used to increase intelligibility of the captured speech. It generally makes audio more pleasant to listen to and can be used to remove irrelevant components from the captured audio. Depending on the use case, background noise reduction may only remove continuous noise such as air-conditioner humming or traffic noise, or it may even aim at removing everything but speech from the captured audio.

Nokia Immersive Voice technology uses machine learning-based, intelligent noise reduction to ensure that only the essential audio is transmitted. The technology adapts to

© 2024 Nokia

the background environment and offers also controls for adjusting the level of noise reduction dynamically.

Another major type of noise that is often causing problems for outdoor audio capture is wind. Immersive Voice's wind noise reduction algorithm can maintain the spatiality of the captured audio and directions of the sound sources while greatly reducing the wind-based disturbances during calls.

Bitrates

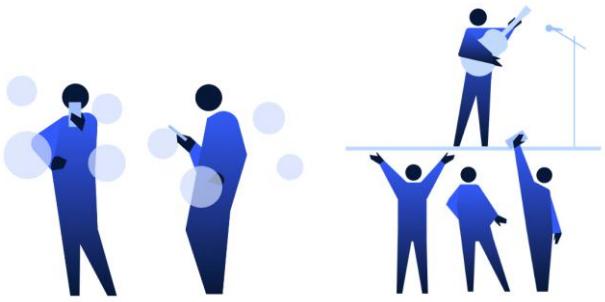
Service providers want to support as many users as possible on their networks in varying network conditions, so the codec will need to be scalable to different bitrates. Nokia Immersive Voice solution is designed to be compatible with the voice-over-IP services in the market and it supports existing and widely used variable-bitrate audio codecs.

Although the requirement of two channels of audio is more than a mono audio transmission, the benefits of a spatial audio experience make it easily justifiable. Additionally, the 3GPP IVAS standard supports a wide range of bitrates. For immersive modes, the IVAS codec operates flexibly between 13.2 and 512 kbit/s, which allows scalable services both in congested networks and high-quality streaming.

© 2024 Nokia

5 A world of new use cases for Nokia Immersive Voice

Nokia Immersive Voice technology unlocks diverse possibilities across various use case categories – enabling Spatial Audio Communication experiences for consumers, enterprises, and industries.



Consumer use cases

Individuals can enjoy closer engagement in interactions with friends and family. Experiences may include:

- Sharing sounds from a local environment, whether recorded, live-streamed, or through a call.
- Using Immersive Voice in social media, sharing immersive stories and live captures.
- Participating in multi-party calls with audio coming from distinct directions. Such examples are voice chats, gaming and social communication.
- Full immersion in metaverse experiences, where audio and visual elements are synchronized.

Enterprise use cases

Immersive Voice enables new functionalities and experiences in enterprise settings. These include:

- Enabling remote customer service with directional audio in one-to-one calls.
- Facilitating team calls that give employees a feeling of proximity to one another.
- Transforming the way teams collaborate enhancing group dynamics and decision-making efficiency.

© 2024 Nokia

Industrial use cases

In industrial settings Nokia Immersive Voice technology contributes to increased safety, security and effective remote monitoring:

- Making one-to-one calls for remote technical assistance with the feel of presence.
- Supporting multi-party calls for remote control and supervision to elevate efficiency.
- Utilizing audio for analytics that contribute to automated industrial processes, such as predictive maintenance.

© 2024 Nokia

6 Conclusion

As we look forward to the future, Nokia anticipates that voice-based user behavior continuing to evolve. Beyond traditional calls, Spatial Audio Communication will expand to include semi-synchronous messaging through popular apps, people sending voice clips to each other, and more extensive use of group calls. With the rise of eXtended Reality (XR) devices and services across industries we believe the scope of voice communication is set to become even broader – with immersion as the defining feature.

Nokia is actively shaping this future landscape through our ongoing work with Immersive Voice technology and audio standardization.

© 2024 Nokia

About Nokia

At Nokia, we create technology that helps the world act together.

As a B2B technology innovation leader, we are pioneering networks that sense, think and act by leveraging our work across mobile, fixed and cloud networks. In addition, we create value with intellectual property and long-term research, led by the award-winning Nokia Bell Labs.

With truly open architectures that seamlessly integrate into any ecosystem, our high-performance networks create new opportunities for monetization and scale. Service providers, enterprises and partners worldwide trust Nokia to deliver secure, reliable and sustainable networks today – and work with us to create the digital services and applications of the future.

© 2024 Nokia

Nokia OYJ
Karakaari 7
02610 Espoo
Finland
Tel. +358 (0) 10 44 88 000
Document code: CID 214046