ANALOG DEVICES
AHEAD OF WHAT'S POSSIBLE™

Future of Immersive Audio: Sound Reproduction Assisted by Computer Vision

Santosh Singh, Senior Engineer, Consumer System Applications, and
Aravind Navada, Director, Consumer System Applications Asia

Abstract

Consumer entertainment is increasingly demanding immersive experiences that allow users to consume content in a way that's indistinguishable from physical reality. Sound plays a key role for immersion in these experiences. Analog Devices envisions a future where audio systems relying on vision intelligence will enable novel sound reproduction techniques that are developed with an improved understanding about how our minds process and localize sounds. Cutting-edge time of flight (ToF) imagers and best-in-class DSPs provide the right mix of technologies and form the ideal platform to realize these next generations of immersive audio systems.

A description of any new age consumer entertainment device reckons the mentioning of the word immersive, but what does it truly mean? In the 1999 hit movie "Matrix," Morpheus asks Neo if what he can smell, taste, or touch is real and then goes on to show that the construct of reality he had known was just about fooling these human senses with computerization. This is what true immersion means and is the goal for any artificial immersive experiences.

To believe that you're truly immersed in an experience, sound and how you experience it are some of the most important parts of the whole thrill. Sound starts a primal reaction in the brain and paves the way on to how we first react to any situation. The brain utilizes sound to form a clearer picture of the environment or situation it is in. Sound plays a key role in delivering the intended immersion by convincing the brain into believing an artificially created immersive experience.

Over the years, sound reproduction technology has made great leaps, evolving from a basic monoaural audio system—that is, single audio channel to today's surround sound systems, which range from a minimal of 5.1 (6-channel) or 7.1 (8-channel) setups suited for home theatres to large 64-channel and higher setups for cinema screens. The spatial sense to sound and its accuracy in these systems is however limited to the number of speakers and the positions they are in.

The next generation of immersive audio systems will be built with novel techniques applied to sound reproduction that will make use of an improved understanding of how our minds process and localize sound. These systems will bring a 360-degree immersive sound experience to home theatres without requiring a large number of speakers positioned around a listener. However, a lack of environmental awareness about the listener and the listening environment will be a major impediment to these systems in meeting the demands of immersive audio. A fusion of vision intelligence with sound reproduction is key in overcoming the impediments and bringing the next generation of immersive audio systems to life.

When we hear sound naturally in a real scenario, our brain is deriving spatial cues about the source of the sound based on only two audio signals—the signals reaching the left and right ears. This is very similar to how our binocular vision system works where a sense of depth is generated inside our brains through a comparison of what the left and right eyes see. Our brains are processing the sound reaching the left and right ears and through a comparison of the amplitude and time delays approximating the location of the sound source. This developed through the evolutionary process as it was critical for survival in the early ages.

Binaural audio reproduction is a technique that aims to replicate the natural listening experience through novel signal processing to generate the same left and right audio signals at each ear as in a real scenario (Figure 1). Achieving this in practice, though, is a tall order and there are various problems that arise.

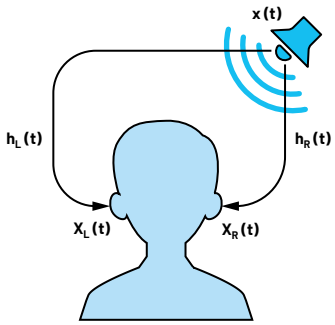

Figure 1. A natural listening scenario from a sound source x(t), X_L (t) is the audio signal reaching left ear, X_R (t) is the audio signal reaching right ear.

One simple way to record binaural audio is to place two microphones—one in each ear canal of a person in a real environment—and record the sound signals at each ear. This recording is called a binaural recording. This is then reproduced through headphones to the ears of a listener. Does this technique work? In a way, yes if the capture and playback are done for the same person. The key reason why this would not work for a different person is because of the way our mind localizes sounds. The impact of our head/pinna/body on sound leaves a specific signature in the frequency domain to aid the sound localization process in our mind. This signature varies from person to person and is called the head-related transfer function (HRTF). For the binaural technique to really work, the HRTF impact on sound must be accurately recreated at the listener's ear during sound reproduction.

HRTFs need to be measured and personalized for each listener; they cannot be solved by a one-size-fits-all solution. Studies have shown that when people experience audio produced using another person's HRTF, their sound localization ability during that experience is markedly reduced.[1,2,3]

There's an even more significant challenge to do binaural audio over loud speakers. First, you have sound signals from multiple speakers interfering with each other. This is called the crosstalk effect (Figure 2). Second, there is the listening environment, which can contribute uninvited effects to the sound before reaching the listener's ear.

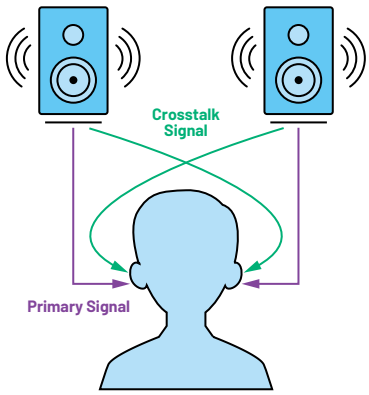

Figure 2. Crosstalk effect in stereo speakers.

Speaker crosstalk, the need for HRTF personalization, and room/listening environment impact are some major impediments in realizing the goal of truly simulating a natural listening experience. A vision system capable of capturing all the details needed about the listener and the listening environment can help in addressing the challenges of binaural audio reproduction.

For example, a camera feeding a computer vision algorithm can be built to capture the three-dimensional structural details of the listening environment (that is, the shape of the listening room, geometric measurement details of different surfaces, and objects present). This insight can be used to compute the impact of the listening environment on sound. Then, appropriate filters and filter coefficients can be used in the sound reproduction system to cancel this uninvited impact. While this type of system is not alien to home theatre audio, it has traditionally relied on the use of a microphone to capture the room's impact on sound during a calibration procedure, which needs to be repeated in case the listening position is changed or there are structural changes to the room.

The vision system can further capture anthropometric measurements such as body position and structural details,[4] which will allow the necessary computations to personalize HRTF for rendering accurate spatial cues (Figure 3). Using information on the listener's head position with reference to the speaker and the head size, a crosstalk cancellation algorithm can be deployed to render real-time binaural audio from the loud speaker setup. This allows the listener to move around without compromising the ideal sound experience (Figure 4).

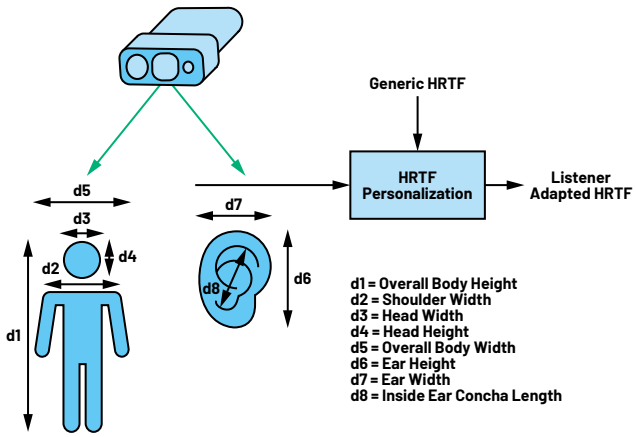

Figure 3. HRTF personalization through anthropometric measurements.

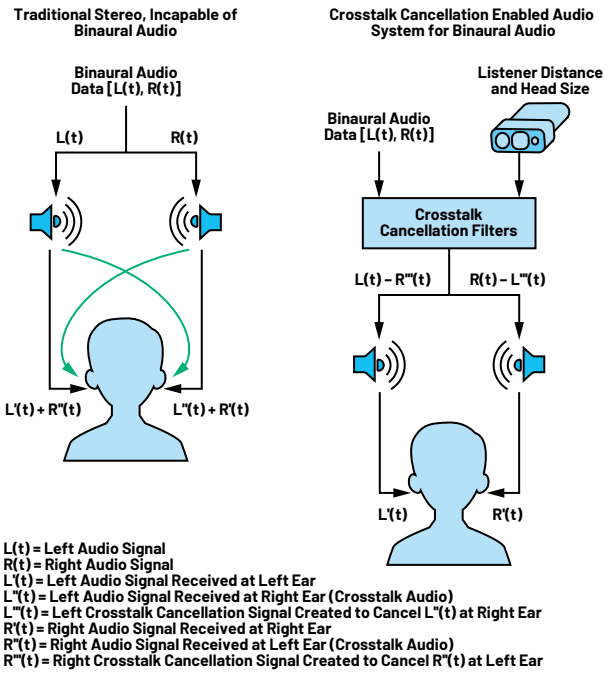

Figure 4. Implementing crosstalk cancellation to enable binaural audio reproduction through free field speaker systems.

A common problem associated with using vision systems is the compromise of user privacy. Processing vision data analytics on the edge using a dedicated compute processor preserves user privacy as the captured camera feed from the vision system is processed in real time and does not need to be stored or transferred to another remote machine.

ADI's latest multicore SHARC® DSPs and cutting-edge ToF imagers deliver the key ingredients for the hardware platform needed to realize the fusion of vision and audio to create the next generation of immersive audio systems (Figure 5).

Our ADSP-SC598 SOCs with dual SHARC cores and an A55 Arm® core supported by a large on-chip memory and DDR interface for external memory is an ideal platform to address the low latency and memory intensive compute requirements for true immersive audio (Figure 6). The compute resources on the SHARC DSP such as ADSP-SC598 enable partitioning of workloads related to audio decode on one DSP core, whereas postprocessing and personalization for audio playback can be implemented on the second SHARC core. The Arm A55 can be employed for a wide variety of control processing.[6] The vision system as described in Figure 5 can be realized with a combination of RGB and depth cameras or a standalone depth camera. Our high resolution 1 MP ToF depth imager ADSD3100 captures depth maps with millimeter resolution and is designed to work across different lighting conditions. This provides highly accurate geometric measurements needed for personalization algorithms previously described (crosstalk cancellation, room equalization, HRTF personalization, etc.).

The ADTF3175 is a 1 MP, 75 × 75 degree field of view (FOV) ToF module based on the ADSD3100 ToF depth imager. It integrates the lens and optical bandpass filter for the imager, an infrared illumination source containing optics, laser diode, laser diode driver and photodetector, a flash memory, and power regulators to generate local supply voltages. The module is fully calibrated at multiple range and resolution modes. To complete the depth sensing system, the raw image data from the ADTF3175 is to be processed externally by the host system processor or depth ISP. The ADTF3175 image data output interfaces electrically to the host system over a 4-lane mobile industry processor interface (MIPI), camera serial interface 2 (CSI-2) transmitter interface. The module programming and operation are controlled through 4-wire SPI and I²C serial interfaces.

Our currently available EVAL-MELODY-8/9 development platform board, EV-2159X/SC59x-EZKIT boards and CrossCore® Embedded Studio (an eclipse-based editor tool) get you up and running with our ADSP SOCs to real-time deploy and debug applications.[7]

The Melody platform is ADI's full signal chain solution for the AVR and soundbar application space. It combines best-in-class ADI components in video, DSP, audio, power, and software into a combined system solution that lets customers get to market quickly with the latest technologies to hit their yearly upgrade windows.[8]

The ToF module, ADTF3175, can be connected to a vision compute platform and connected over to a Melody board to build the hardware platform for the next-gen immersive audio system (Figure 7). An RGB camera can be coupled to the ADTF3175 ToF module to form an RGBD camera for enhanced vision analytics.

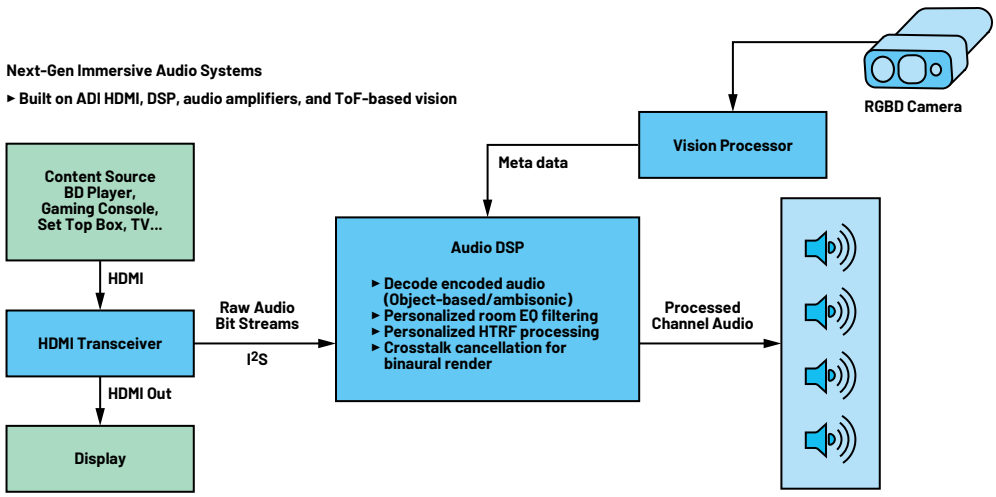

Figure 5. Next-generation immersive audio systems.

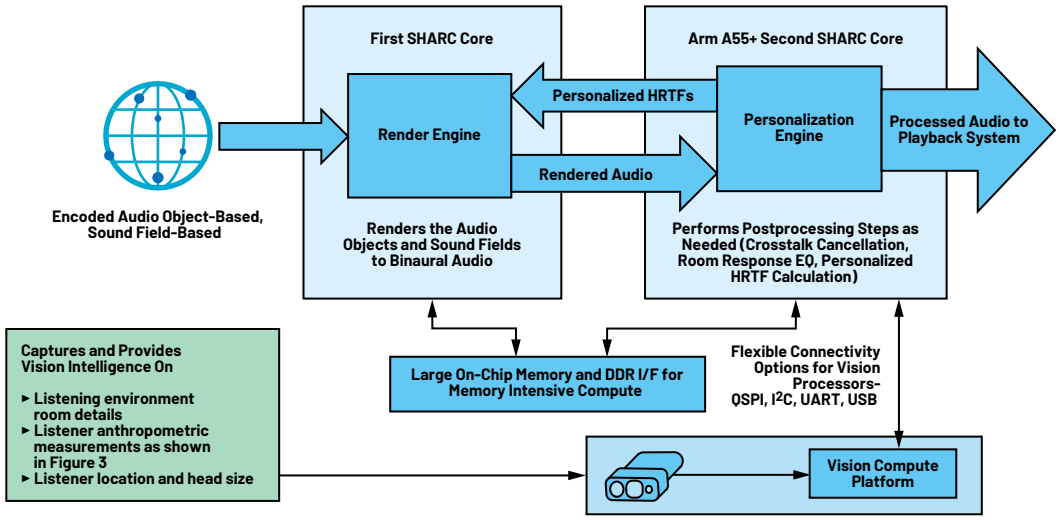

Figure 6. System partitioning of next-generation immersive audio systems.

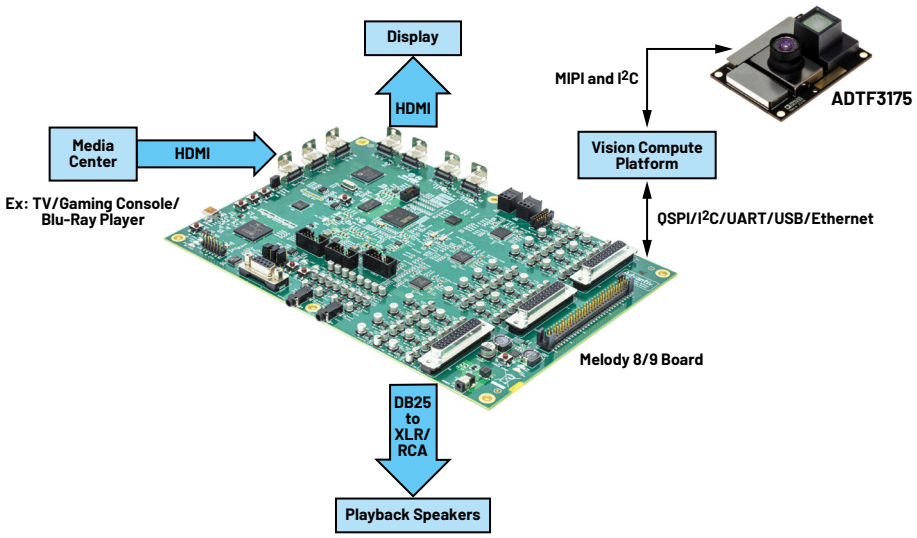

Figure 7. Realizing immersive audio systems with ADI platforms.

Conclusion

With our portfolio of solutions across DSPs, HDMI transceivers, Class-D amplifiers, and ToF imagers, ADI is committed to the pursuit for true immersive audio systems that aim to reproduce sounds indistinguishable from sounds in the real world.

References

[1] Philipp Paukner, Martin Rothbucher, and Klaus Diepold. "Sound Localization Performance Comparison of Different HRTF-Individualization Methods." Technische Universität München, April 2014.

[2] Parham Mokhtari, Ryouichi Nishimura, and Hironori Takemoto. "Toward HRTF Personalization: An Auditory-Perceptual Evaluation of Simulated and Measured HRTFs." International Conference on Auditory Display (ICAD), July 2008.

[3] Jenny Claudia and Christoph Reuter. "Usability of Individualized Head-Related Transfer Functions in Virtual Reality: Empirical Study With Perceptual Attributes in Sagittal Plane Sound Localization." JMIR Serious Games, September 2020.

[4] Geon Woo Lee and Hong Kook Kim. "Personalized HRTF Modeling Based on Deep Neural Network Using Anthropometric Measurements and Images of the Ear." Applied Sciences, November 2018.

[5] Sanket Nayak and Mitesh Moonat. "EE-436: Using ADSP-SC59x/2159x High Performance FIR/IIR Accelerators." Analog Devices Inc., June 2022.

[6] "ADSP-SC59x/2159x SHARC Series Doubles Portfolio Performance, Enabling Platforms for Complex Audio Applications." Analog Devices, Inc.

[7] "CrossCore® Embedded Studio." Analog Devices, Inc.

[8] "Home Theater and Gaming." Analog Devices, Inc.

About the Authors

Santosh Singh graduated with a bachelor's degree in electronics and communications in 2016 from the Birla Institute of Technology and Science. He's been working with Analog Devices for the past six years, initially as a college intern and then a full-time employee. He started his career as a digital designer working on various cutting-edge technology products for the consumer market. Santosh is currently a senior system applications engineer with the Consumer Business unit. His main focus is on solving key system challenges that help bridge the gap between applications and technologies.

Aravind K. Navada manages the Consumer System Applications team with particular focus on customers across Asia. He has more than 20 years of experience in IC design with a history of bringing a variety of mixed-signal IoT and consumer SOCs to the market. He is currently focused on developing platforms and solutions for the consumer market that intend to maximize Analog Devices' product insertion. Aravind has a bachelor's degree in electronics and communication from University of Visvesvaraya College of Engineering (UVCE), Bangalore and a master's degree in microelectronics from the University of Texas at Dallas.