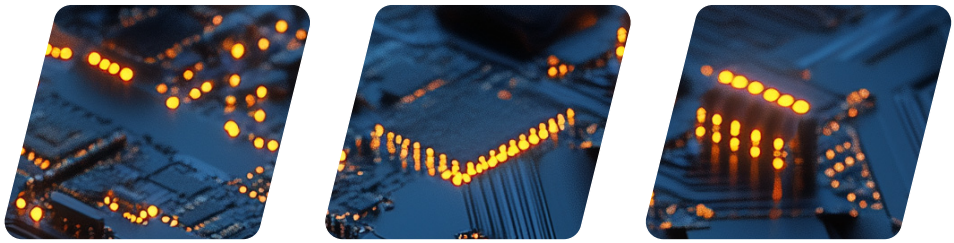
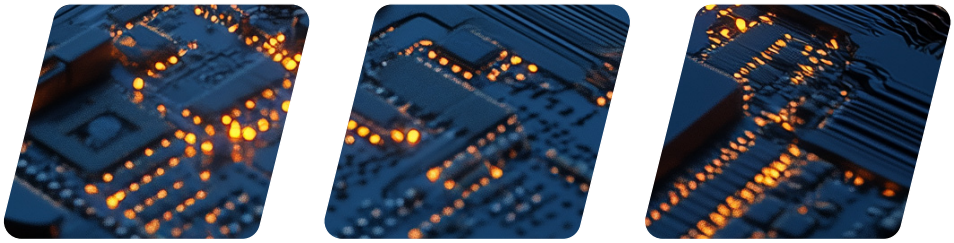


The Acoustic Modality in Autonomous Robotics



White Paper

Authors:

Deepak Boppana, Senior Director, Segment Marketing, Lattice Semiconductor
Jay Shu, Director, Systems Engineering, Analog Devices, Inc.

DISCLAIMERS

Lattice makes no warranty, representation, or guarantee regarding the accuracy of information contained in this document or the suitability of its products for any particular purpose. All information herein is provided AS IS, with all faults, and all associated risk is the responsibility entirely of the Buyer. The information provided herein is for informational purposes only and may contain technical inaccuracies or omissions, and may be otherwise rendered inaccurate for many reasons, and Lattice assumes no obligation to update or otherwise correct or revise this information. Products sold by Lattice have been subject to limited testing and it is the Buyer's responsibility to independently determine the suitability of any products and to test and verify the same. Lattice products and services are not designed, manufactured, or tested for use in life or safety critical systems, hazardous environments, or any other environments requiring fail-safe performance, including any application in which the failure of the product or service could lead to death, personal injury, severe property damage or environmental harm (collectively, "high-risk uses"). Further, buyer must take prudent steps to protect against product and service failures, including providing appropriate redundancies, fail-safe features, and/or shut-down mechanisms. Lattice expressly disclaims any express or implied warranty of fitness of the products or services for high-risk uses. The information provided in this document is proprietary to Lattice Semiconductor, and Lattice reserves the right to make any changes to the information in this document or to any products at any time without notice.

INCLUSIVE LANGUAGE

This document was created consistent with Lattice Semiconductor's inclusive language policy. In some cases, the language in underlying tools and other items may not yet have been updated. Please refer to Lattice's inclusive language FAQ 6878 for a cross reference of terms. Note in some cases such as register names and state names it has been necessary to continue to utilize older terminology for compatibility.

ABSTRACT

As autonomous robotics and multimodal Vision-Language-Action (VLA) models evolve, acoustic perception remains a critically underutilized modality. This white paper outlines a high-performance, edge-compute architecture for real-time acoustic spatial tracking and signal isolation. By combining Analog Devices, Inc. (ADI) A²B® (Automotive Audio Bus) Link technology for physical array distribution with the Lattice Semiconductor Holoscan Sensor Bridge solution and NVIDIA® GPU-accelerated Delay-and-Sum (DAS) beamforming, robotic platforms can achieve deterministic, low-latency auditory awareness. This pipeline produces unclipped Pulse-Code Modulation (PCM) audio that is designed to support seamless integration with downstream audio-processing workflows for robotics.

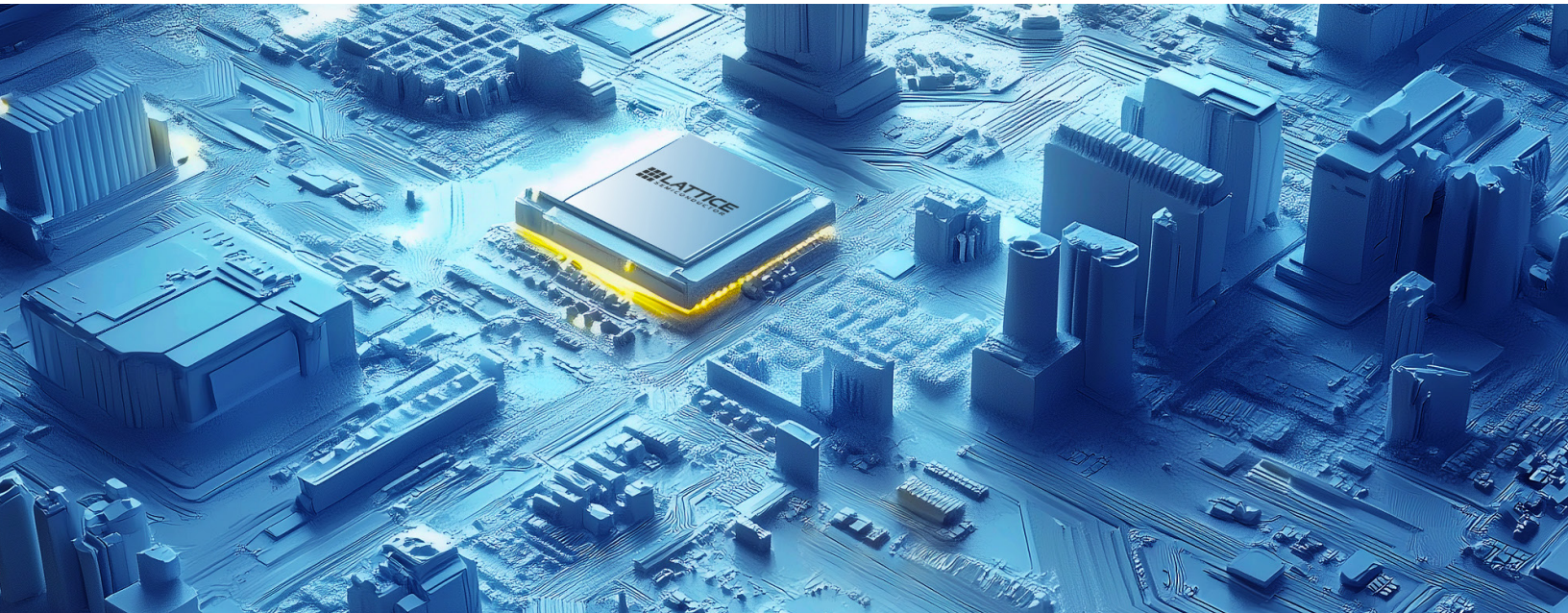


TABLE OF CONTENTS

The Missing Modality: Human-Robot Interaction via Acoustic Perception

Modern autonomous systems rely heavily on sensors like lidar, IMUs, and computer vision frameworks to navigate and map their environments. However, as robots transition from isolated industrial tasks to collaborative, human-centric environments, vision alone becomes insufficient. Vision is inherently directional and is easily hindered by occlusion, poor lighting, and narrow fields of view.

To achieve true collaborative autonomy, robots require seamless Human-Robot Interaction (HRI). Acoustic perception provides a 360°, non-line-of-sight sensor modality that allows a robot to perceive and respond to human operators regardless of where its cameras are currently pointing.

The Necessity of the Microphone Array

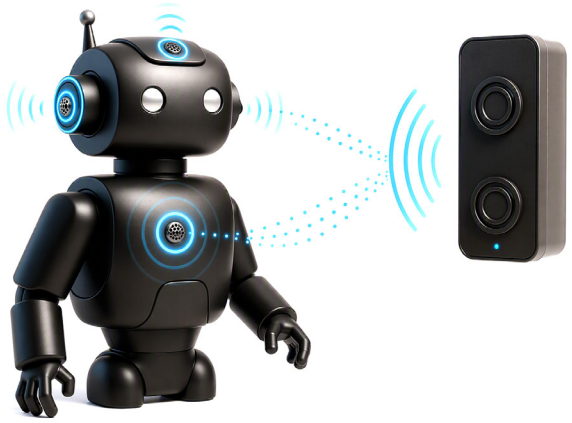
Simply attaching a single microphone to a robot chassis is inadequate. A single acoustic sensor captures a flattened, indistinguishable mix of human speech, structural motor rumble, and ambient room noise. To make acoustic data actionable for higher-level decision-making, a spatial array is necessary.

Incorporating a distributed microphone array into the robotics stack enables two critical HRI capabilities:

- Spatial Attention (The “Where”): By calculating the Time Difference of Arrival (TDOA) across the array, the robot can determine the approximate 3D spatial coordinates of a speaker. This auditory anchor allows the system to autonomously orient its chassis or direct its primary optical sensors toward the human operator.
- Acoustic Isolation (The “What”): Using Delay-and-Sum beamforming, the array acts as a highly directional, steerable acoustic lens. It physically extracts the human operator’s vocal commands while simultaneously rejecting off-axis machinery and environmental noise.

By leveraging the array to isolate clean, localized speech, the robot can reliably feed clean acoustic prompts into downstream Voice-to-Text engines or Vision-Language-Action models, enabling natural, hands-free collaborative control. However, physically integrating such an array across a robotic chassis introduces significant hardware challenges. See Figure 1.

Figure 1: Distributed microphone array placement across a robotic chassis for 360-degree acoustic perception.



The Physical Layer: Distributing Arrays with ADI A²B

For spatial tracking to be accurate, microphones must be physically spaced apart. A wider baseline yields higher angular resolution. However, wiring multiple individual analog microphones across a robot's chassis introduces physical cabling issues, cable weight, electromagnetic interference (EMI), and complex clock-synchronization challenges.

The A²B Solution and TDM Scalability

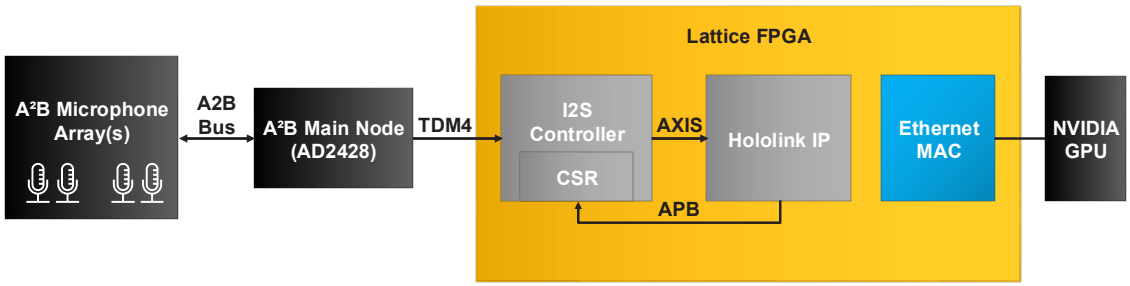
ADI's A²B transceivers form the ideal backbone for robotic microphone arrays.

- Daisy-Chained Simplicity: A²B distributes high-fidelity digital audio, I2C control, and power over a single unshielded twisted-pair (UTP) wire. Microphones can be daisy-chained across the robot's head or chassis, drastically reducing wiring harness weight and complexity.
- Deterministic Synchronicity: Beamforming requires absolute phase alignment. A²B provides strict, deterministic latency, guaranteeing that the audio sampled at the robot's left ear is designed to be perfectly time-aligned with the right ear before it ever reaches the compute module.
- Time Division Multiplexing (TDM) Signaling: To scale beyond standard 4-channel configurations, the architecture utilizes TDM signaling. TDM enables efficient aggregation of high-channel-count arrays, easily supporting up to 32 microphones on the bus. This TDM formatting is currently implemented and supported directly by the Lattice FPGA interface, allowing engineers to scale the acoustic resolution of the robot without requiring complete hardware or compute redesigns.

FPGA Implementation: A²B I2S Ingress and Deterministic Audio Transport

The design operates across two asynchronous clock domains: the I2S bit-clock (BCLK) domain driven by the A²B transceiver, and the system clock domain of the sensor-bridge interconnect. All clock-domain crossing and data-synchronization logic is handled internally by the I2S controller, enabling seamless operation across a wide range of BCLK frequencies without requiring hardware redesign. See Figure 2.

Figure 2: End-to-end audio data path from A²B microphones through Lattice FPGA-based I2S Controller and Hololink AXIS streaming to an NVIDIA GPU over Ethernet.



The Inter-IC Sound (I2S) controller is fully software-configurable and supports:

- Multiple I2S stream formats, including Mono, TDM2, TDM4, and higher-channel configurations
- Configurable I2S audio bit depths
- Configurable audio data block size

The Lattice FPGA captures multichannel audio data directly from the I2S interface and converts it to Hololink IP's AXIS data stream for transfer to the host system over Ethernet. Audio samples are efficiently packaged into clean, preconfigured data blocks and transferred across clock domains through a design intended to preserve data integrity and provide predictable latency, ensuring reliable real-time operation. This flexible, software-defined architecture supports a wide range of audio formats and channel configurations without hardware changes and integrates naturally with GPU-accelerated beamforming and spatial-audio pipelines. By keeping time-critical data capture and transport in hardware while offloading compute-intensive audio processing to the GPU, the system achieves high performance, low latency, and an optimal balance between scalability and efficiency.

The Compute Layer: GPU Acceleration

Processing high channel count audio requires evaluating over 500,000 audio samples per second. Traditionally, executing real-time audio spatial processing and beamforming requires integrating a dedicated Digital Signal Processor (DSP) into the hardware design to address both latency and computational demands.

Zero-Copy Execution

NVIDIA Holoscan shifts the sequential CPU loops to a set of parallel GPU operations. Once the raw audio enters the pipeline via the Lattice Holoscan Sensor Bridge, it remains in GPU VRAM. TDOA calculations, cross-covariance matrices, and Least Squares spatial triangulations are executed across parallel thread blocks, completely avoiding CPU context-switching overhead and ensuring strict timing deadlines are met.

Acoustic Isolation: The GPU Beamformer

Knowing where a sound originates addresses only part of the robotics audio processing chain. To feed a clean prompt to a downstream model, the audio processing chain must physically tune out the ambient room noise using a true physics-based Steered DAS beamformer.

- **Time-Alignment:** Raw buffers are shifted backwards in time by the exact fractional lags calculated during the TDOA phase.
- **Coherent Summation:** The aligned signals are summed. Because the target voice is perfectly aligned, it adds linearly. The incoherent ambient noise adds quadratically, resulting in an improvement to the Signal-to-Noise Ratio (SNR).
- **Finite Impulse Response (FIR) Convolution:** The spatially isolated beam is passed through a hardware-accelerated FIR bandpass filter utilizing GPU constant memory. This final stage efficiently wipes out structural low-frequency rumble and high-frequency hiss without relying on AI-based noise suppression that might distort vocal clarity.

Bridging the Gap: Audio into VLA/VLM Models

The ultimate objective of this pipeline is to provide actionable intelligence to the robot's higher-level decision-making software.

Data Formatting for Neural Networks

Standard audio processing often applies digital gain or dynamic range compression optimized for human listening. In contrast, this pipeline is engineered to output clean, unprocessed PCM waveforms. By preserving the exact physical acoustic amplitude and dynamic range of the environment, the beamformer output is designed specifically to feed directly into Automatic Speech Recognition (ASR) nodes in the audio processing path. Maintaining this absolute volume reference provides critical spatial and vocal effort context before the data is passed to Vision-Language-Action and Vision-Language Models (VLM).

The Transport Layer

Since the acoustic pipeline operates independently on the GPU, it avoids bottlenecking the CPU. The resulting 1D audio tensors and 3D spatial coordinates can be passed along to downstream processing tasks on the GPU. This allows the acoustic tracking vector to be natively published in the GPU alongside other telemetry, providing immediate, synchronized auditory context to complement the robot's visual and navigational stack. By publishing the acoustic tracking vector and isolated PCM stream natively within the GPU environment, developers can effortlessly bind audio data to frameworks like ROS 2 or Isaac ROS. These frameworks are designed to eliminate the traditional latency penalties of routing audio through discrete Advanced Linux Sound Architecture (ALSA) subsystems and CPU-bound middleware, with the architecture intended to align the acoustic modality with visual frames and inertial telemetry at the fusion layer.

Conclusion: Unifying Multimodal Perception

As the robotics industry accelerates toward autonomous, collaborative systems, the ability to perceive and isolate human intent in noisy, unstructured environments is no longer optional. Vision-Language-Action models represent a massive leap in robotic intelligence, but they are fundamentally constrained by the quality and dimensionality of their sensory input.

The architecture outlined in this paper solves the historical bottlenecks of edge-compute acoustic processing. By leveraging ADI A²B technology, the physical burden of array distribution is reduced to a single lightweight cable. Through the Lattice Holoscan Sensor Bridge, deterministic data ingress is guaranteed without requiring complex, custom hardware redesigns. Finally, offloading the Delay-and-Sum beamforming and spatial tracking directly to the NVIDIA GPU enables execution with zero-copy efficiency.

Ultimately, this integrated approach transforms audio from a passive, ambient signal into a deterministic, highly spatialized sensor stream. When fused alongside high-resolution vision and precision inertial data on the same edge-compute architecture, this acoustic pipeline unlocks the next generation of truly collaborative, multimodal autonomous robotics.

READY TO LEARN MORE?

To evaluate the FPGA solution, visit the Lattice Holoscan Bridge Sensor Interfaces Reference Design webpage or contact your Lattice representative. For more information on ADI A²B technology, visit the A²B Automotive Audio Bus Technology webpage or contact holo.scan@analog.com.

TECHNICAL SUPPORT ASSISTANCE

Submit a technical support case through **www.latticesemi.com/techsupport**.
For frequently asked questions, please refer to the Lattice Answer Database at **www.latticesemi.com/Support/AnswerDatabase**.

V.1 WP0055