Seeing with Sound: AI-Based Detection of Participants in Automotive Environment from Passive Audio

J. Sieracki, R. Yang, A. Vamos, M. Caggiano

AIoT Center of Excellence Engineering Team, Renesas Electronics

Abstract

Existing ADAS solutions for car environmental awareness (cameras, LiDAR, ultrasonic, etc.) require targets to be in a clear line of sight from the sensor. The target must be illuminated by some source of energy, so systems are affected by dust, weather, lighting, and obstacles. We address those limitations using a passive acoustic solution that “listens” to the environment. It can hear potential targets around corners or out of sight over a distance, providing early warning that supplements and improves other ADAS systems. We aim to detect a variety of road participants including sirens, approaching vehicles, bicycles, and even pedestrians. We discuss use cases and challenges, present an inexpensive reference architecture based on automotive grade components, and report on the updated state of development with initial validation results.

Introduction

Existing solutions for car environmental awareness typically use cameras or active sensing (e.g., lidar, radar, ultrasonic sonar). These approaches can be costly and have significant limitations. All require targets to be in a clear line of sight from the sensor. They require the target to be specifically illuminated by light or some other source of energy and are affected by dust, weather, and obstacles. [1]

The Renesas passive acoustic solution [2] complements existing methods by addressing all these limitations. Audio does not require a clear line of sight, and, in the case of a loud target such as an emergency vehicle siren, can work over long distances. Over moderate distances, targets can be tracked and located. Passive audio does not require specific illumination of the target with lighting or other energy sources. It can work even when targets are obscured or around corners.

We have demonstrated that audio can also be used to make class distinctions about target types, having developed models that can detect and classify emergency vehicle sirens, automobiles, motorbikes, bicycles, and pedestrians (joggers). Audio, coupled with a suitable AI model, provides a powerful capability that complements and improves on other solutions. It can supplement camera-based or lidar-based recognition as well as radar or ultrasonic with non-line-of-sight sensing at a low-cost point.

Providing a new modality that can “see” around obstacles and corners to produce an early warning – even when a target is out of sight of the driver or camera element – presents a powerful addition to safety and awareness.

This paper updates a report presented at the AES International Conference on Automotive Audio in 2022 [3].

Challenges

Our goal is to detect the full diversity of road participant targets and localize bearing-to-target using ambient sound, captured using a MEMS microphone array and processed on a standard, low-cost automotive microcontroller.

Specific challenges for this solution hinge on its use of passive monitoring. Since the system relies on sound emitted by targets rather than on illuminating targets with a known energy source, we must account for a wide range of variation in each target type. We will also be limited by the relative sound levels of the target compared to the background noise levels.

A related challenge for use of audio sensors is their placement on the vehicle. The vehicle itself shadows the sound field, so a small array placed on top of the vehicle might provide 360-degree coverage but would be less effective for nearby targets that are below roof-level. A set of arrays placed on the perimeter of the car (e.g., roof corners, bumpers, mirrors, corner lighting positions, etc.) would provide better coverage of proximity sounds and a wider baseline for localization. Our goal is to support both short- and wide-baseline options with a general reference design, giving each manufacturer flexibility to meet their design and performance requirements. Therefore, our system is designed around sub-arrays of microphones that can be placed at arbitrary, fixed distances from one another on the vehicle, enabling different manufacturers to deploy different configurations.

Additional challenges are presented by virtue of the automotive application. The physical environment is challenging and requires automotive-qualified hardware components, including microphones, processing and networking devices. Besides the increased operating temperature range, the long lifetime of cars must be supported by all components. Water and dust pose extra challenges for the microphone module design. Automotive qualification of hardware components demonstrates such durability by testing under AEC-Q100 / AEC-Q200 specifications. [2] These define failure tests though extreme environment cycles, mechanical, electrical, and other stresses, to guarantee very low defect-rates over projected part lifetimes of 15 years or more.

Automotive software must also be reviewed and qualified for safety assurance reasons. Finally, the system itself needs to be minimal so as not to add excess weight, take too much volume, or require substantial wiring during vehicle manufacture. All these challenges are addressed by our system.

Target Use Cases

Fig. 1 Typical use case is nearby targets in slow speed, obstructed view situations. Sirens will be detected at greater range under wider driving conditions. Image is copyright © Renesas Electronics, used with permission.

The Seeing-with-Sound (SWS) system is designed with both human oriented advanced driver assistance systems (ADAS) and autonomous driver stacks in mind. Its purpose is to augment other sensor systems to (a) confirm targets in line of sight, while (b) extending the augmented awareness around corners where existing systems fail.

Our road participant targets currently include other motor vehicles, bicycles, joggers, and emergency vehicle sirens. Additional targets may be added in the future.

Fig. 2 Initial data was collected on compact sedan ego vehicles (e.g. Toyota Corolla.) The system was tested in a variety of actual neighborhoods, complex obstacle areas and purpose-built closed courses.

The primary use case for SWS is in low-speed, obstructed view situations, as depicted in *Figure 1*. *Figure 2* shows other test locations used in development. For example, at intersections, entrances, alleys, etc., where the ego vehicle is either stopped or moving slowly. Passive audio is ultimately limited in range by both the loudness of the target and the loudness of the environment. This signal to noise ratio constrains us for most targets to conditions under which the ego vehicle self-noise and wind-noise does not overwhelm the sound emissions of the target.

Fig. 3 We are also working with an autonomous trucking company to instrument heavy vehicles in live tests ranging from engine idling to highway speeds. Arrows show test array locations. Inserts (top right) show exterior automotive rated microphone modules arrays which can be integrated behind bodywork or fairings in production.

An exception, of course, is emergency vehicle sirens. Since these are designed to be heard at a distance, the reliable range of operation will be much larger. Noise encountered while traveling at speed, however, is still a challenge. *Figure 3* shows a Volvo semi tractor-trailer instrumented for live testing. Data was acquired

using this vehicle from idle up to full highway speeds. With this heavy vehicle, self-noise with engine idling is already 83 dB(A) at the roof. Together, wind and engine noise on the highway reach 105 dB(A).

We target siren detection out to 1 km under road speed driving conditions or slower driving conditions in cluttered building areas. Under highway noise conditions, we are targeting siren detection on 300 m. We target other moving cars and motor vehicles out to 50 m line of sight and up to 30 m around blind corners. Bicycles and pedestrians are targeted at closer range, with an emphasis on detecting them in obstructed view situations where other sensors fail.

System Architecture

Microphones

The system is designed around small microphone subarrays of two or more inexpensive MEMS mics placed around a vehicle. This facilitates incorporation by a designer into their choice of body locations, so long as the positions are not substantially obstructed from ambient sound.

Figure 4 illustrates positions that have been tested for a sedan. These include the corners of the car (near head- and taillights or in bumpers), on the corners of the roof, and in the center areas of the roof. Other locations are possible, including center bumpers, mirrors, and so forth.

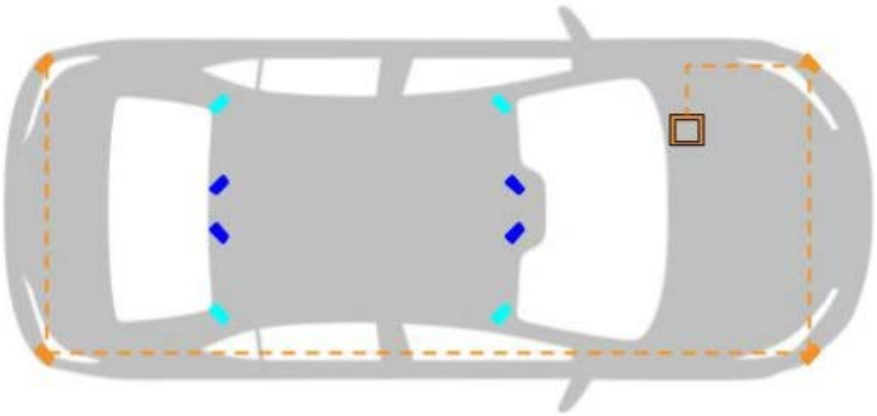

Fig. 4 360˚ coverage is achieved with small mic arrays around the car, at any locations convenient to the designer, such as car corners or roof (e.g., orange, blue, cyan bars). A2B uses a single, lightweight cable bus routed to an MCU (yellow square), reducing complexity and weight significantly from traditional analog audio wiring.

Exterior microphones in an automotive setting must deal with challenging acoustic and environmental conditions. This includes sound levels of 130dB SPL or more, along with wide ranges of dust, moisture, and temperature. We use microphone array packages [5] that are automotive grade and rated IP6K9K for exterior weather conditions. With protected sound ports and gaskets around connectors, driving tests even under wet conditions are possible. Microphone packages are shown in *Figure 3* inserts.

Audio Signal Bus

These pulse density modulation MEMS microphones are coupled in subarrays to Analog Devices (ADI) A2B® bus slave nodes, and over that bus to a master node for audio processing. The bus [6] provides us with 32-bit, 48 kHz sample rate, phase coherent digital channels from each microphone to our processing

point. It does so with a lightweight, daisy-chained twisted-pair cable which saves significant weight and complexity from using traditional point-to-point audio wiring.

While we do not need 48 kHz for this application, we do make use of the bit depth. It is extremely important that the system have sufficient signal head room so that quiet target signals can be reliably discovered in loud environments.

Processing

The A2B bus terminates in a master node, coupled to an automotive grade MCU [7]. As discussed further in the next section, our algorithms are designed to be lightweight and not require expensive DSP or Neural co-processors.

In our reference implementation, one MCU is sufficient to handle acquisition and down-sampling of the audio, all machine learning inference, localization processing, and outbound communications for four mic subarrays. That is enough to cover the entire 360 degrees around the vehicle.

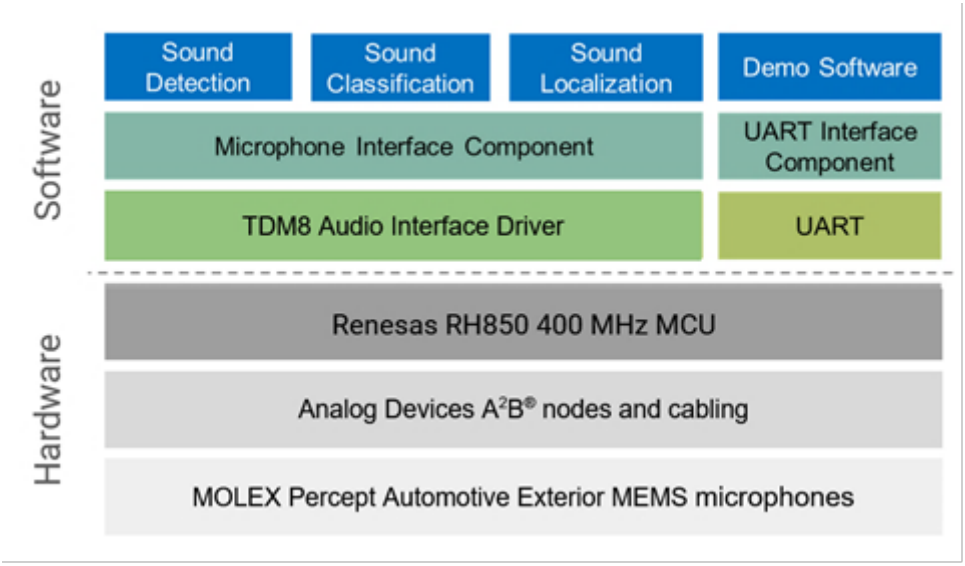

Fig. 5 Reference design HW/SW stack.

Figure 5 illustrates the software / hardware stack employed in our reference design. Because the software is lightweight, alternatives to the single MCU solution include splitting processing of sub-arrays to even less expensive local MCUs or incorporating portions of it into an autonomous driving stack.

The code is fast and compact, and Reality AI machine learning inference does not require any special computing hardware. In its current implementation, we run audio input processing, AI target detection, and angle of arrival (AOA) on one core with code including the A2B stack occupying ~300KB. This is still operating in debug and not optimized for production. On the RH850 MCU, this leaves significant CPU, Flash and RAM for other tasks. We can operate either 4 microphone (180-degree coverage) or 8 (360-degree coverage) microphone configurations.

Theory of operation

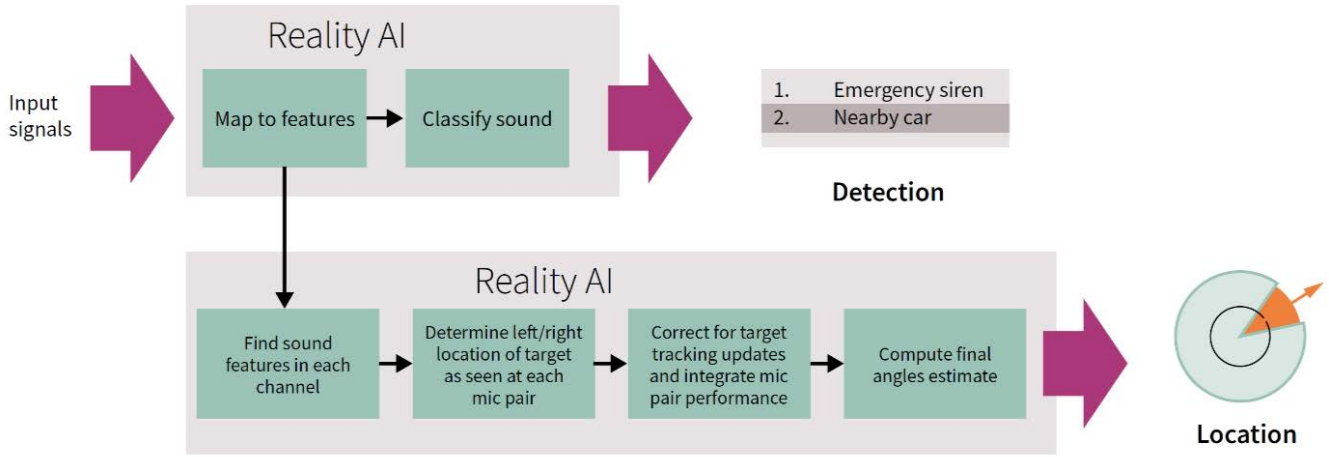

Fig. 6 Processing flow diagram for Renesas Reality AI SWS system

As illustrated in *Figure 6*, our solution comprises two systems working together: a classifier that detects and distinguishes targets (e.g., sirens, cars) and a localizer that identifies the angle of arrival (AOA) of sound from the detected target.

The machine learning (ML) portions of the system were developed using Reality AI Tools™, a feature-mapping and ML development system designed specifically to address complex signal problems. This technology creates compact embedded inference code modules that have been used in the reference implementation.

Target detection and classification

Input from each microphone channel is processed to form a collection of time-frequency feature components. These components are evaluated with machine learning inference code to discriminate targets from background noise and to distinguish targets. To aid in suppressing false positives, agreement on target presence and type is required among multiple microphone channels. Additional time smoothing methods are employed to suppress brief errors due to sound field dynamics.

This subsystem outputs the presence or absence of each class of target. For ADAS display purposes, certain targets may be given priority or the system can easily be configured in a way to report multiple target types simultaneously.

Target Localization

AOA determination of the sound from a target can be based on two factors. The first is the relative loudness of sounds arriving at different points on the vehicle. The second is phase differences in the arrival of sounds in local subarrays. Localization is based on signal processing techniques supplemented with machine learning.

The primary basis for determining the AOA in this demonstration is based on phase delay rather than loudness comparisons. This dictates that at least two microphones work together in a fixed distance

relationship to determine angle up to 180 degrees accuracy. As discussed, for flexibility, our algorithms are built around subarray modules of two or more microphones that can be placed arbitrarily around the vehicle to detect 360 degrees of sound.

To reduce the computational requirements for AOA signal processing, our algorithm leverages information about the target and its component time-frequency features from the machine learning inference layer. Rather than relying broadly on cross correlation and coherence as is commonly done to blindly identify targets in noise, we instead focus on specific spectral elements associated with the detected target and use those to analyse the relevant phase delay between microphones. The phase delay provides an estimated AOA for each microphone subarray.

Information from each microphone subarray determined to be in the direct path of the sound source (i.e., not blocked by the body of the vehicle) is combined to get a final estimate of the AOA relative to the vehicle.

Target tracking and time smoothing

The current system evaluates the sound field at a rate of up to 12 times per second, determining the likelihood of a target in the sound field. For each such target, it estimates a target bearing. This information is combined to smooth the instantaneous output which both reduces false positives and suppresses noise in the AOA computation. Once a target is identified as moving in a particular angular direction (e.g., right to left or left to right), further tracking filters are applied to predict and test AOA changes over time.

Performance

The system is in development for a variety of road participant targets, including emergency vehicle sirens, motorized vehicles, bicycles, and joggers. *Table 1* shows our target specifications for the first two groups. Goals for bicycles and joggers are not yet public. Here we report test results for sirens; information on other targets will be available later this year.

	Detection	**Bearing**
Emergency Vehicle Sirens	95% within 1 km (moderate noise) 500 m (city noise) 300 m (highway noise)	95% within ±22.5˚ (clear) ±45˚ (corner)
Motorized vehicles (including EV)	95% within 50 m (line of site) 30 m (around corner)	

Table 1 Target specs for emergency sirens and on-coming cars. Moderate noise is 62 dB(A) background, typical of suburban setting with nearby roads, city noise is 85 dB(A), typical of mid-day city traffic. We are still determining criteria for highway noise levels of 95+ dB(A).

The system was initially developed using extensive data collection in a purpose-built set of alleys and intersections laid out in a large, paved lot. We are continuing development and testing of the system now in actual streets and intersections. Live test statistics are determined by comparing system output to ground-

truth video cameras monitoring the scene. We use the percentage of time in which correct predictions are made as the accuracy metric.

We aim for accurate classifiers in realistic background noise levels that serve our practical use cases. 62 dB(A) is typical of sound levels measured in a suburban or office park settings with faster roads in the distance. 85 dB(A) is typical of city noise with busy day traffic. [8]

As discussed, we have also started testing using a semi-truck in noise conditions ranging from engine-idle noise at 75 dB(A) to highway noise at 95 dB(A). In these practical tests, we discover that not only does the loudness of the noise increase with a heavy vehicle but its character also changes. The engine noise is much broader spectrum and both engine and wind noise mask siren frequencies much more than was observed on the sedan.

Figure 7 shows current performance curves for sirens on our sedan test vehicle. Our reference level is a typical US ambulance siren, measured at 102 dB(A), 20 meters from the vehicle. These curves are based on live tests using quieter sirens to simulate distance out to ~ 1km with larger distances simulated using well known equations for sound pressure reduction with distance. Recordings were remixed with live background noise recordings to maintain target levels. Accuracy is measured as the proportion of time we correctly report the presence of siren in our output display when the siren is operating. We meet our target 1 km 95% detection rate for an emergency vehicle siren even in the “around corner” condition with 65 dB(A) noise.

“Free field” corresponds to a condition in which the siren is line-of-site from our test vehicle. This is relatively uncommon in actual traffic beyond a few hundred meters but we include it here for comparison. “Around corners” corresponds to a condition with the test vehicle positioned at a 4-way intersection and the siren around one or more corners to the left or right of the vehicle. The second condition is a typical non-line-of-site situation with occlusions and reflections in which the emergency vehicle has a high likelihood of becoming a driving safety factor.

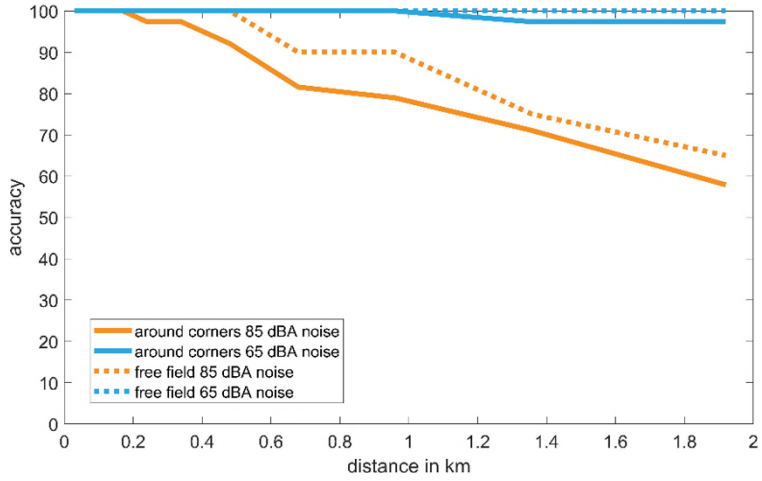

Fig. 7 Siren detection performance in two background noise levels on sedan ego-vehicle, for both line-of-site and non-line-of-site tests.

Figure 8 shows performance on a semi-truck ego-vehicle at idle, low speed and highway speeds. We believe emergency vehicle detection should also remain practical even at highway noise levels and fast ego car speeds, provided exterior wind noise issues are properly mitigated. This is a subject of ongoing effort. The current heavy vehicle highway tests indicate about a 300 m range of detection for a siren approaching from behind.

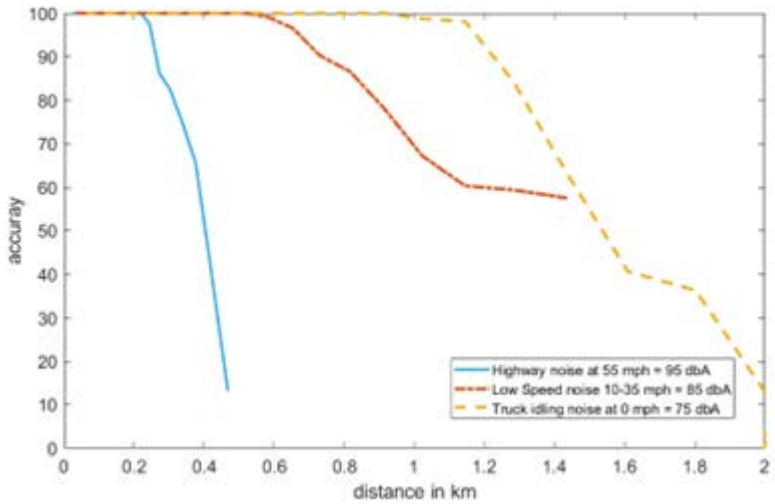

Fig. 8 Performance on semi-truck ego-vehicle at idle, low speed, and highway speeds.

Audio engineers may be surprised at the ±22.5˚ bearing accuracy target, however, we have found that directional accuracy for nearby moving targets has important practical limits. While several well-known signal processing methods can provide tight estimates of the direction of a stationary sound source, challenges here are twofold: (1) Vehicles emit signature sounds from many locations (tires, engine, exhaust, etc.) (2) targets move over the period of the analysis. So, for targets even at moderate distance and speed, the angular span of the sound source plus the travel distance makes for a significant margin of error. Sirens add the additional difficulty that they are loud and focal, thus prone to issues of echo and multi-path fade when not in line-of-site.

We therefore specify our bearing accuracy with this in mind, understanding that the important practical point of our SWS system is to provide an early warning and a general direction to the driver (or autopilot) sufficient to act on.

Finally, we compared microphone array locations, including bumper corners, roof corners, roof center and side mirrors. We found no significant difference between locations in detection accuracy. Performance for AOA was also similar across locations in free field, however, around corners the bumper corner location had up to a 10% time-correct advantage, when judged at ±22.5˚ bearing accuracy. This is likely due to the bumper's closer proximity to the corner, giving it a "field of view" advantage. We conclude, however, that all locations are viable.

Conclusions

We have described ongoing development work for an automotive-qualified, ADAS system that "listens" to passive sounds in the vehicle environment and provides early warning of presence and direction of other

road participants. It can hear potential targets around corners or out of sight over a distance, providing early warning that supplements and improves other ADAS systems.

Our functionality covers important use cases including early warning of obstructed view emergency vehicles and those approaching from behind on a highway, as well as low speed driver assistance with avoiding obstructed view cars and other nearby road participant collision threats.

We will continue to develop the system to better perform at highway speeds and on loud, heavy vehicles. This will be addressed by our partners with physical improvements such streamlined fairings and body panel microphone placements, as well as by our own team with ongoing development in signal processing and increasingly noise robust AI detection.

The system is built around low-cost, automotive-qualified hardware that will be available for third party testing this year. Initial release will cover emergency vehicle sirens, followed by updates for cars and other motor vehicles. Several highly challenging additional target cases, including bicycles and joggers, are in development for future release.

References

1. G. Smith, *Types of ADAS Sensos in Use Today*, dewesoft.com (2021) https://dewesoft.com/daq/types-of-adas-sensors
2. Reality AI® Seeing-with-Sound™ system: https://reality.ai/automotive-sound-recognition-localization/
3. J Sieracki, M. Boehm, P. Patki, M. Caggiano, M. Noll, *Seeing with Sound: Detection and localization of moving road participants with AI-based audio processing,* AES Conference on Automotive Audio, Dearborn, Mi, USA, June8-10, 2022.
4. AEC-Q100 and AEC-Q200 requirements http://www.aecouncil.com/AECDocuments.html
5. Molex® Precept automotive MEMS microphones packages: https://www.molex.com/en-us/products/sensors/percept-sensors
6. Analog Devices (ADI) A²B® Audio Bus: https://www.analog.com/en/applications/technology/a2b-audio-bus.html
7. Renesas® RH850 Automotive MCUs: https://www.renesas.com/us/en/products/microcontrollers-microprocessors/rh850-automotive-mcus#featured_products
8. J. Rodrique., *The Geography of Transport Systems*, Fifth Edition, Routledge (2020) https://transportgeography.org/contents/chapter4/transportation-and-environment/noise-levels/

(Rev.1.0 Mar 2020)

Corporate Headquarters

TOYOSU FORESIA, 3-2-24 Toyosu, Koto-ku, Tokyo 135-0061, Japan
https://www.renesas.com

Contact Information

For further information on a product, technology, the most up-to-date version of a document, or your nearest sales office, please visit:
https://www.renesas.com/contact-us

Trademarks

Doc Number: R33WP0005EU0100