Human Systems Engineering and Design (IHSED 2026), Vol. 237, 2026, 873–880
https://doi.org/10.54941/ahfe1008225

Event-Level Acoustic Emergency Detection to Support Operator Awareness in Public Transport

Diva Pant[1] and Franziska Schmalfuss[2]

[1]Department of Sound, Audio and Voice Functions, IAV GmbH, Munich, Germany
[2]Department of Intelligent Perception Functions, IAV GmbH, Stollberg, Germany

ABSTRACT

We built a non-visual emergency sound detection system for public transport cabins that reports event level alerts (no faces, no identities). The system consists of two stages. First, Anomalous Sound Detection (ASD) learns what normal cabin audio sounds like (e.g., bus, tram, metro, car) and identifies audio segments that deviate from this background. In this context, an anomaly corresponds to sounds such as human distress, crying, glass breaking, or sirens that stand out from engine noise and passenger chatter. ASD scans continuous audio and flags candidate segments using an anomaly score derived from reconstruction error and a threshold calibrated on normal data. Second, Sound Event Detection (SED) processes only these flagged segments and assigns an emergency related label (human distress, crying, glass break, siren). The system was evaluated on synthetic cabin audio scenes that reflect real cabin conditions, including polyphonic overlaps and SNR levels. The proposed models outperform standard baseline models. For ASD, a CNN autoencoder achieved AUC 96.98% and F1 91.71%, compared to 91.82% AUC and 81.87% F1 for a fully connected autoencoder. For SED, EfficientSED exceeded CRNN baselines (F1 59.55% vs. 50.49%, ER 0.75 vs. 0.94, PSDS1 0.40 vs. 0.33). The research results show that acoustic centric AI can provide timely and stable emergency cues under realistic public transport noise and overlap conditions, supporting operator monitoring without visual sensing.

Keywords: Acoustic event detection, Emergency monitoring, Public transport, Human in the loop systems, Operator support

INTRODUCTION

Human Factors Engineering involves understanding the need for comprehensive integration of human capabilities (Ahram & Karwowski, 2009) (cognitive, physical, sensory, and team dynamics) into a system design, beginning with conceptualization and continuing through system disposal. The primary concern for human factors in engineering is the need to effectively integrate human capabilities with system interfaces to achieve optimal total system performance (use, operation, maintenance, support, and sustainment). Human factors engineering utilizes comprehensive task analyses to help define system functions and then allocates those functions to meet system requirements. The goal of HSI is to optimize total

Received April 22, 2026 Revised June 13, 2026 Accepted July 7, 2026 Available online September 25, 2026

system performance, accommodating the characteristics of the user population that will operate, maintain, and support the system, and minimize life-cycle costs (Folds et al., 2008). HSI experts work within the Systems Engineering (SE) process to ensure that all human considerations are integrated throughout system design, development, fielding, sustainment, and retirement. The attention to human systems integration in system development programs drove hundreds of human-centered design improvements. Efforts were concentrated to maximize total system performance through improvements in human workload, ease of maintenance, and personnel safety which resulted in a cost avoidance of billions of dollars and prevention of hundreds of fatalities and disabling injuries for the system (Booher and Minninger, 2003). Highly automated vehicles (AVs) operating at SAE Level 4 and beyond fundamentally change the role of humans inside the vehicle cabin, as the absence of a vehicle in-charge shifts responsibility for safety and security entirely to onboard systems (e.g., Ajith & Sharmila, 2024, Tsiktsiris, Lalas, Dasygenis, Votis & Tzovaras 2020; Tsiktsiris, Lalas, Dasygenis & Votis, 2025). Research indicates that passengers' negative attitudes toward automated public vehicles are linked to safety, security concerns, and the absence of driver support during emergencies or incidents (Kassen-Noors et al., 2021; Schmalfuß et al., 2026). In such vehicles, in-cabin monitoring becomes a critical requirement to ensure the safety of occupants, protect the vehicle from misuse or damage, and enable reliable real-time supervision in the absence of human authority (e.g., Mishra et al., 2021, Mishra et al., 2020). From a human factors perspective, safety in highly automated vehicles extends beyond crash avoidance or functional reliability and includes occupants' physical protection, comfort, and perceived security during travel. Prior work distinguishes between occupant safety, referring to protection from physical harm and discomfort, and occupant security, referring to protection against deliberate harmful actions such as vandalism, abuse, or harassment (e.g., Mishra et al., 2022; Tsiktsiris, Lalas, Dasygenis & Votis, 2025).

Most in-cabin monitoring research relies on visual sensing for tasks like analysing passenger behaviour and detecting abnormal events (e.g., Ajith & Sharmila, 2024; Tsiktsiris, Lalas, Dasygenis & Votis, 2025). Trajectory-based methods underperform with restricted camera views due to extensive preprocessing requirements, such as tracking and occlusion detection, needed to filter out false data. Supporting multiple camera types, especially wide-angle lenses, is a key constraint. While pose estimation works well with standard lenses, it faces core challenges when used with top-down camera angles (Tsiktsiris et al., 2020). Real-time in-cabin perception, visual abnormal event detection remains challenging due to strong context dependence, occlusions, and the difficulty of inferring urgency or intent from visual information alone. Other approaches need more attention. Nandi et al. (2020) developed a pressure-sensitive seat design to identify passenger states, but it does not apply to standing passengers. Prior work highlights that audio information can capture events that are not visually apparent, thereby enriching situation awareness in complex environments (e.g. Tsiktsiris, Lalas, Dasygenis & Votis, 2024). The integration of audio sensing aligns with principles of multimodal perception and redundancy, supporting more robust detection of safety- and security-relevant events in real time. Audio signals provide a complementary modality for understanding in-cabin situations that may not be fully observable visually or by analyzing seat pressure. Acoustic events such as raised voices, sudden impacts, breaking objects, or alarms can indicate abnormal or safety-critical situations even when no clear visual cues are present. Tsiktsiris et al. (2024) described a multi-model approach, including RGB video frames, depth data and audio data to detect bag snatching, falling, fighting, vandalism and normal behavior. Thus, the stand-alone effect of audio data analysis has not been investigated.

Despite the growing body of work on in-cabin monitoring, audio-based anomaly and event detection has received comparatively limited attention, and its systematic integration into intelligent in-cabin monitoring systems for highly automated vehicles remains an open research challenge. This gap is particularly relevant for public and shared mobility services, where incidents such as passenger aggression, vandalism, or distress have been identified as significant concerns in the absence of an onboard authority figure. As automated public transport systems evolve, the role of remote operators becomes increasingly complex, requiring continuous monitoring across multiple vehicles and dynamic environments. Research in human-automation interaction emphasizes that operators benefit from multimodal cues that provide early indicators of abnormal system states, particularly when visual information is incomplete, occluded, or delayed (Folds et al., 2008; Booher & Minninger, 2003). Audio cues can reveal events that are not immediately visible on camera—such as distress vocalizations, sudden impacts, or escalating conflict—and therefore serve as an essential complement to visual monitoring. This aligns with foundational principles of Human Systems Integration, which advocate designing systems that augment human perceptual strengths while compensating for limitations in high-demand operational contexts (Chapanis, 1996). Integrating acoustic anomaly detection into the monitoring workflow thus enhances operator situational awareness, reduces cognitive load, and contributes to a more resilient and human-centered safety architecture for automated public transport.

Motivation

A central motivation for this system is the need to reduce operator cognitive load. Control room personnel often monitor dozens of video streams simultaneously, a task known to degrade attention and increase the likelihood of missed events. Acoustic alerts act as attention-directing cues, enabling operators to shift focus only when necessary. This selective engagement supports more efficient monitoring and reduces fatigue.

Privacy considerations also shape the system's design. Unlike visual surveillance, which captures identifiable features, the acoustic module processes short, audio segments and outputs only event-level labels. No biometric or personal information is stored or inferred, making the system suitable for public transport environments where ethical and regulatory constraints limit intrusive sensing.

The system supports human-in-the-loop decision making by providing interpretable alerts that operators can verify through visual feeds or contextual information, rather than acting autonomously. Thresholds for anomaly detection and event classification are transparent and adjustable, enabling operators to calibrate sensitivity according to operational requirements. This approach strengthens trust and ensures the system functions as an assistive tool rather than an opaque decision-maker.

SYSTEM OVERVIEW

The system architecture integrates acoustic monitoring into the broader safety workflow of public transport and automated vehicle cabins. Audio captured through onboard microphones is continuously processed by the emergency detection module, which identifies anomalous or safety-critical events and forwards only relevant alerts to a remote operator control room. This human-centered design prevents operators from being overwhelmed by raw audio streams or continuous surveillance data. Instead, the system acts as an intelligent filter that surfaces only high-value cues requiring human attention.

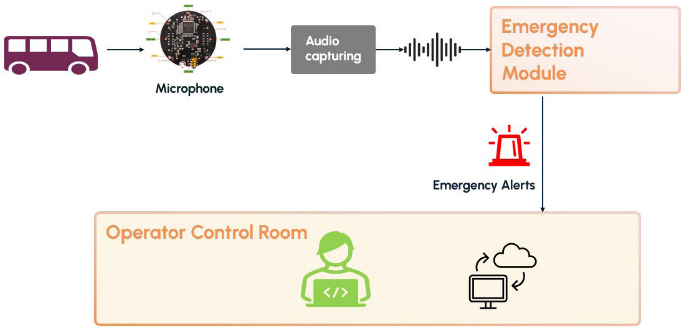

Figure 1: Emergency detection module to assist operator monitoring for safety critical situation.

Building on this conceptual workflow, the system adopts a two-stage architecture inspired by human auditory perception. The first stage, Anomalous Sound Detection (ASD), learns the acoustic signature of normal cabin environments and flags deviations. The second stage, Sound Event Detection (SED), classifies these anomalous segments into interpretable categories such as human distress, crying, glass breakage, or sirens. Together, these components reduce false alarms, prioritize high-value events, and provide operators with actionable, semantically meaningful information. The modularity of the architecture ensures compatibility with existing monitoring systems without requiring intrusive sensing or identity-based data. Figure 1 gives a high-level emergency detection system overview.

METHODS AND RESULTS

The system uses synthetic transport soundscapes to train emergency-focused acoustic models, enabling controlled and realistic evaluation without recording real emergencies. This approach allows the creation of diverse, repeatable, and ethically sourced training data that reflects the acoustic complexity of public transport cabins. Background recordings from buses, trams, metros, and cars were curated as individual audio files representing typical operational conditions, including engine noise, ventilation systems, door mechanisms, and varying levels of passenger chatter. These recordings capture the natural variability of cabin acoustics, such as reverberation patterns, fluctuating noise floors, and intermittent mechanical sounds that characterize real-world environments. Emergency-indicative foreground events—such as crying, distress vocalizations, glass breaking, and sirens—were sourced as separate audio files for each class and overlaid onto the background recordings at multiple signal-to-noise ratio levels and temporal positions. This approach follows established practices in synthetic soundscape generation for sound event detection and enables the creation of controlled yet realistic polyphonic scenes (Mesaros et al., 2021; Gharib et al., 2019).

To ensure methodological transparency and operational realism, audio segments were standardized to a 16 kHz sampling rate and generated in short windows of approximately 1–2 seconds to support rapid operator alerting. This segment length reflects the temporal granularity at which human operators typically perceive and react to auditory cues in control room settings. Polyphonic scenes were created by allowing multiple foreground events to overlap within a single segment, simulating realistic scenarios such as distress vocalizations occurring simultaneously with mechanical noise or conversations. Overlapping events also reflect the masking effects common in public transport cabins, where high-energy noise sources can obscure subtle but safety-critical cues. This file-based mixing strategy ensured that the dataset captured the acoustic complexity, unpredictability, and contextual variability characteristic of real transport environments. These methodological choices were guided by human-factors requirements: the need for rapid detectability, robustness to noise, and alignment with operator workflows in control rooms where early auditory cues can significantly reduce cognitive load and response time.

Beyond the construction of the dataset, the system design was guided by two core principles: human-centered realism and scalability. Human-centered realism ensured that the audio scenes reflected the perceptual challenges faced by operators monitoring multiple vehicles simultaneously. This included simulating conditions where important events may be partially masked, occur abruptly, or arise in the presence of competing noise sources. Scalability ensured that the system could be extended to additional event classes or deployed across different vehicle types without requiring a complete redesign of the processing pipeline. The two-stage architecture supports this modularity. The Anomalous Sound Detection (ASD) module was implemented using a convolutional autoencoder trained exclusively on normal cabin acoustics. By learning the statistical structure of typical background noise, the autoencoder identifies deviations that may indicate safety-critical events. This approach aligns with anomaly detection principles used in safety-critical monitoring systems, where deviations from learned norms often serve as early indicators of abnormal system states.

The Sound Event Detection (SED) module builds on the ASD output by classifying anomalous segments into interpretable categories such as human distress, crying, glass breakage, or sirens. A lightweight transformer-enhanced architecture was selected to balance performance with computational efficiency, enabling real-time inference on embedded hardware. Transformers have demonstrated strong performance in modeling temporal dependencies in audio, particularly in polyphonic environments where overlapping events must be disentangled. This architectural choice reflects human-factors requirements such as low latency, interpretability, and robustness to noise. The modular separation between ASD and SED also supports operator trust by providing a clear, interpretable pipeline: first detect anomalies, then classify them into meaningful categories.

Evaluation demonstrated strong performance across both modules. The convolutional autoencoder achieved high discrimination between normal and anomalous audio, enabling reliable filtering of continuous streams and reducing the number of segments requiring classification. This is particularly important in operator-in-the-loop systems, where excessive false alarms can lead to alarm fatigue and reduced vigilance. The SED module performed robustly for high-energy events such as sirens and glass breakage, outperforming conventional CRNN baselines. These events exhibit distinct spectral signatures that remain detectable even under moderate noise conditions. Low-energy vocalizations such as crying were more challenging, consistent with prior findings on variability in human vocal distress and the difficulty of detecting subtle vocal cues in noisy environments (Khandelwal et al., 2024). Nevertheless, the system demonstrated resilience to noise and polyphonic overlap, supporting deployment in real transport cabins where acoustic conditions are unpredictable. In comparison to single-stage detection approaches, the proposed two-stage pipeline shows more stable behavior in continuous cabin audio streams. The anomaly detection stage limits the propagation of routine background activity, ensuring that the event classifier operates primarily on acoustically relevant segments. This selective processing reduces unnecessary activations and leads to more consistent detection behavior under varying noise levels and overlapping sound conditions. While quantitative evaluation indicates improved detection quality, the more relevant observation is the consistency of performance across acoustically challenging scenarios. The system maintains reliable detection even when foreground events are partially masked or occur alongside competing noise sources, which is a common characteristic of real public transport environments. This behavior is enabled by the separation of detection and classification tasks. The anomaly detection module focuses on modeling normal acoustic patterns, while the event detection module specializes in distinguishing between emergency-relevant sounds. This division reduces model confusion in complex scenes and allows each module to operate within a more constrained and stable decision space.

Overall, the results suggest that structuring acoustic detection as a staged and selective process provides advantages not only in detection performance but also in the reliability and interpretability of system outputs, which are critical for operator-facing monitoring systems. From a human-systems integration perspective, the system has been designed with future verification and validation needs in mind. A key principle in HSI is that evaluation must consider not only algorithmic accuracy but also how effectively a system supports human operators in real monitoring environments (Booher, 2003). Factors such as trust calibration, workload, and situational awareness will need to be assessed during subsequent system-level testing rather than isolated usability studies. The architecture of the acoustic monitoring module supports these goals by providing interpretable alerts, adjustable sensitivity thresholds, and a modular processing pipeline that can be integrated into existing control room workflows. By surfacing only relevant acoustic events and filtering out routine cabin noise, the system is intended to reduce the cognitive burden associated with continuous visual monitoring and enable more efficient allocation of operator attention. These design considerations lay the groundwork for future empirical studies examining how acoustic cues can enhance human performance and contribute to a more resilient, human-centered safety architecture for automated public transport.

CONCLUSION

This work presents a human-centered acoustic monitoring approach designed to enhance safety and situational awareness in public transport and automated vehicle cabins. By combining anomaly detection with targeted event classification, the system introduces a privacy-preserving and interpretable layer of acoustic analysis that complements existing visual monitoring solutions. Its modular architecture and adjustable sensitivity thresholds are intended to support operator workflows by surfacing potentially relevant acoustic events without exposing raw audio or identifiable information.

From a human-systems integration perspective, the system has been developed with operator needs and future evaluation requirements in mind. The design emphasizes transparency, interpretability, and alignment with ethical principles governing public surveillance. Rather than replacing visual monitoring, the acoustic module is conceived as an assistive tool that may augment operator perception, particularly in situations where visual cues are limited, occluded, or delayed. These design considerations lay the foundation for future empirical studies examining how acoustic cues could support trust calibration, workload management, and situational awareness in remote supervision contexts.

Several challenges remain. Low-energy vocalizations such as crying are inherently difficult to detect in high-noise environments, and the use of synthetic soundscapes—although realistic—cannot fully capture the variability of real-world acoustic conditions. Future work will incorporate field recordings, adaptive thresholding strategies, and multimodal fusion to improve robustness and ecological validity. Additional research will also explore operator-in-the-loop evaluations to assess how acoustic alerts influence attention allocation, decision-making, and overall monitoring performance.

Overall, this work illustrates the potential of intelligent acoustic technologies to contribute to human-centered safety architectures for automated mobility systems. By prioritizing interpretability, privacy, and operator support, the system lays important groundwork for creating safer, more responsive, and more trustworthy public transport environments.

ACKNOWLEDGMENT

This work is part of the OptiPEx project (No. 101146513), funded by the European Union. The views expressed are solely those of the authors and do not represent the European Union or CINEA. The EU and the granting authority are not responsible for these views. Portions of this manuscript were assisted by a large language model (Microsoft Copilot) for text restructuring and language refinement. All content was reviewed and verified by the authors.

REFERENCES

Ahram, T. and Karwowski, W. (2009). Human systems integration modeling using systems modeling language. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 53(24), 1849–1853. Sage Publications, Los Angeles, CA.

Ajith, K. and Sharmila, R. (2024). AI-Based In-Cabin Monitoring System for Autonomous Vehicles. IGI Global.

Booher, H. (2003). Handbook of Human Systems Integration. New Jersey: Wiley.

Chapanis, A. (1996). Human Factors in Systems Engineering. Hoboken, NJ: Wiley.

Folds, D., Gardner, D. and Deal, S. (2008). Building Up to the Human Systems Integration Demonstration. INCOSE INSIGHT, 11(2).

Gharib, M., et al. (2019). Anomalous Sound Detection in Real Environments. DCASE Workshop.

Kassens-Noor, E., Cai, M., Kotval-Karamchandani, Z. and Decaminada, T. (2021). Autonomous vehicles and mobility for people with special needs. Transportation Research Part A.

Khandelwal, A., Das, A. and Chng, E. (2024). Challenges in Acoustic Event Detection in Noisy Environments. IEEE Transactions on Affective Computing.

Mesaros, A., Heittola, T., Virtanen, T. and Plumbley, M. (2021). Sound Event Detection: A Review. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Mishra, A., Cha, J. and Kim, S. (2020). In-Cabin Monitoring System for Irregular Situations. ISOCC.

Mishra, A., Kim, J., Kim, D., Cha, J. and Kim, S. (2021). HCI-Based In-Cabin Monitoring System. IHCI.

Mishra, A., Lee, S., Kim, D. and Kim, S. (2022). In-Cabin Monitoring System for Autonomous Vehicles. Sensors.

Nandi, P., Mishra, A., Kedia, P. and Rao, M. (2020). Real-Time Autonomous In-Cabin Sensory System to Detect Passenger Anomaly. IEEE IV.

Schmalfuß, F., Eichstädt, M., Kinnula, A., Blasco-Giménez, E. and Lucas-Alba, A. (2026). Highly Automated Public Transport for Everybody. Humanist Conference.

Tsiktsiris, D., Dimitriou, N., Lalas, A., Dasygenis, M., Votis, K. and Tzovaras, D. (2020). Real-Time Abnormal Event Detection for Autonomous Shuttles. Sensors.

Tsiktsiris, D., Lalas, A., Dasygenis, M. and Votis, K. (2024). Multimodal Abnormal Event Detection in Public Transportation. IEEE Access.

Tsiktsiris, D., Lalas, A., Dasygenis, M. and Votis, K. (2025). A Complete In-Cabin Monitoring Framework for Autonomous Vehicles. IET Intelligent Transport Systems.