Development of an audio event detection system for complex situational awareness for intelligent urban safety control

Youngjun Choi
Convergence Technology Research Institute
AIBLab Inc.
Seongnam, South Korea
yjchoi@aiblab.co.kr

Sungsik Park
Convergence Technology Research Institute
AIBLab Inc.
Seongnam, South Korea
sungsikp@aiblab.co.kr

Junwook Lee
Convergence Technology Research Institute
AIBLab Inc.
Seongnam, South Korea
junux@aiblab.co.kr

***Abstract* — In smart city environments, ensuring citizen safety requires intelligent monitoring systems capable of rapidly detecting unexpected incidents. Conventional video-based surveillance systems face inherent limitations such as restricted visibility and privacy concerns. This study proposes an audio event detection system for urban safety control, designed to identify critical events such as screams, vehicle horns, sirens, dog barks, and explosions. The system transforms acoustic signals into Mel-spectrograms and employs pretrained neural network models to extract features and perform multi-class classification. Detected events are then mapped to risk levels and integrated with a complex situational awareness service. Experimental results demonstrate that the proposed system achieves strong performance on both public datasets and real-world urban recordings, thereby validating the potential of audio-based real-time threat detection technologies.**

Keywords*— AIoT, *smart city, edge computing, audio event detection, audio surveillance, AI detection

I. INTRODUCTION

In recent years, the application of artificial intelligence (AI) technologies has been increasingly emphasized to enable the rapid detection of and efficient response to diverse urban hazards. Urban environments are constantly exposed to risks such as traffic accidents, fires, crimes, and large-scale disasters, all of which require real-time detection and timely intervention to ensure citizen safety. However, traditional video-based surveillance systems remain constrained by several critical limitations, including restricted visibility under low-light or adverse weather conditions, privacy concerns associated with continuous visual monitoring, and the high costs of installation and maintenance. To overcome these limitations, recent studies have shifted toward intelligent monitoring approaches that integrate audio and IoT sensor data to complement vision-based systems. In particular, audio data play a vital role in complex situational awareness, as they directly capture critical signals such as screams, vehicle horns, sirens, and explosions, which are strongly associated with urban safety and emergency events.

To guarantee real-time responsiveness and processing efficiency, reliance solely on cloud servers is insufficient; instead, processing audio event streams with AI models at the edge is increasingly recognized as a practical solution. Edge-based audio processing minimizes data transmission latency, alleviates the computational burden on centralized cloud resources, and enables immediate detection and faster decision-making. Nevertheless, as the number of audio sensors and event streams managed by a single edge device increases, challenges emerge due to limited computational resources that restrict the ability to process all data simultaneously. Furthermore, effective urban safety control requires systems that can flexibly update detection models to address emerging event types and provide remote management capabilities for diverse audio sensors. These challenges underscore the necessity for edge audio detection systems to evolve beyond simple real-time detection toward more advanced architectures that account for scalability, efficient resource management, and long-term system adaptability.

To address these issues, this paper proposes a novel edge audio detection system architecture and AI processing method designed to optimize resource utilization through a distributed and multi-layer structure. The proposed system is capable of efficiently handling the full pipeline of audio data processing, including collection, stream preprocessing, AI-based detection, post-processing of results, and system-level management. By enabling distributed execution of these tasks, the architecture supports flexible deployment of diverse edge-level AI models, while ensuring the reliable detection of critical urban safety events and their seamless integration into complex situational awareness services. Ultimately, this work aims to establish the scalability and cost-effectiveness of AI-based infrastructure for intelligent urban disaster and safety control systems, presenting audio detection at the edge as a key enabler of citizen safety in future smart cities.

II. RELATED WORKD

The intelligent management of urban control increasingly relies on the integration of diverse sensor applications. In smart city environments, heterogeneous data sources such as IoT sensors and audio sensors are used to support multi-dimensional situational awareness in domains including traffic management, public safety, and disaster response. Recent studies emphasize the role of edge computing in minimizing data transmission latency and ensuring real-time analytical performance, thereby addressing the limitations of cloud-centric architectures and enabling time-critical decision-making for urban safety control [1][2].

For efficient situational analysis, it is essential that diverse events can be detected promptly at the edge level. For instance, research in intelligent transportation has explored edge-based detection of traffic incidents and acoustic signals such as honking or collisions for real-time response [3]. Similarly, environmental sensing has been applied to monitor air quality, temperature, humidity, and noise levels to improve citizen safety and well-being [4]. More recently, multimodal approaches that integrate audio with other sensing modalities have gained attention, as they provide richer contextual information and enable robust situational awareness [5].

A common insight across these studies is that intelligent edge-level analysis is indispensable for advancing smart city safety management. In particular, audio event detection has been recognized as a core component of urban control systems, as it directly reflects emergency situations such as accidents, crimes, and disasters [6]. Building on these insights, this paper extends prior work by focusing on an edge-based audio event detection system architecture designed to enhance complex situational awareness for intelligent urban disaster and safety control.

III. EDGE COMPLEX CONTEXT-AWARE SYSTEM ARCHITECTURE

A. Architecture Design

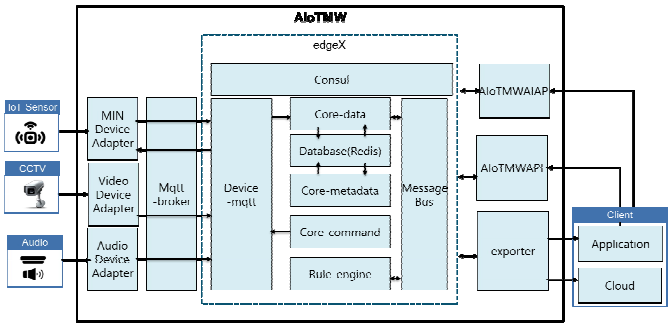

Fig. 1. Edge complex context-aware system architecture

As shown in Figure 1, the system for multimodal data-based situation analysis for efficient city control is divided into a device layer, an edge middleware layer(AIoTMW) , and an application layer. Devices include IoT sensors, CCTV, and audio devices. The edge middleware detects simple situational context events based on data from each device and transmits them to the edge middleware. The edge middleware then has a pipeline that performs complex situational awareness of the multimodal event stream and delivers the results to the cloud or service applications.

B. Components

- Adapter: Detects simple situational events from raw data from each device by deploying and operating various AI models and transmits them to edge middleware.
- edgeX: An open platform for edge data processing, it performs core common functions such as device management, data reception and rule processing, and command processing.
- AIoTMWAPI: An extended service of edgeX, it handles device management, device model management, and interfaces with upper-level applications.
- AIoTAIAPI: For complex situational awareness, edgeX's rule system processes simple situational events and analyzes spatiotemporal patterns in the event stream to process complex situational awareness and deliver it to applications.

C. Data Flow

- Adapter -> MQTT Broker: Each adapter transmits a simple, analyzed situation event message through the broker. The simple situation event contains information about the detection model type and detection details for each device.
- MQTT Broker-> device service -> core-data: Simple situation information is verified through the device service and then loaded through the core-data service.
- core-data -> message bus -> Rule system: New data received is published through the message bus and transmitted to the rule system, where it is processed according to the user context rules defined in the rule system.
- Rule system -> AIoTMWAIAPI : Simple situation events detected in the rule system are analyzed into complex situations based on time patterns within the travel time window through AIoTMWAIAPI.
- AIoTMWAIAPI -> Client : The analyzed complex situation event is propagated to the client.

IV. AUDIO CONTEXT EVENT DETECTION

In urban safety management, audio information provides critical cues that are often difficult to capture through video or environmental sensors alone. Sounds such as screams, horns, and explosions can serve as immediate indicators of incidents, enabling rapid responses within complex situational awareness systems. Moreover, audio data can be

collected at relatively low cost and remain effective even in challenging conditions such as nighttime, adverse weather, or areas not covered by cameras. Thus, audio event detection has emerged as a core component for real-time threat detection and intelligent urban safety control in smart city environments. This section categorizes audio event types, describes the YAMNet model employed for detection, and discusses practical considerations for deployment in real-world urban contexts.

A. Adapter Architecture Description

Figure 2 illustrates the adapter architecture of the proposed urban safety control system. The overall structure is divided into four layers: Device, Adapter, edgeX, and AIoT Service.

The Device layer consists of CCTV cameras, sound detectors that capture audio events, and time-series sensors for environmental data collection. These devices generate raw data in heterogeneous formats and communication protocols, making direct integration into higher-level platforms challenging.

To address this, the Adapter layer serves as an interface between devices and edgeX. The VMS Adapter converts CCTV/VMS video streams into a standardized format compatible with edgeX. Similarly, the Sound Adapter processes audio detection results, and the Sensor Adapter normalizes time-series sensor data. In essence, adapters play the role of standardizing heterogeneous device data into unified event streams for edgeX integration.

In the edgeX layer, Device-mqtt collects the standardized data using the MQTT protocol. Coredata and Redis DB manage data storage, while the Rule Engine (Kuiper) applies user-defined rules to filter and process incoming events.

Finally, the AIoT Service layer incorporates the AIoTAIAPI as the core component for complex situational awareness. A Flink-based stream engine performs spatiotemporal pattern analysis and real-time event stream processing, while the API Server enables integration with upper-level applications. The processed results are distributed to external systems through MQTT and ZMQ protocols.

Thus, the adapter architecture is a key enabler for integrating and standardizing heterogeneous device data at the edge platform, ensuring scalability and reliability for intelligent urban disaster and safety control systems.

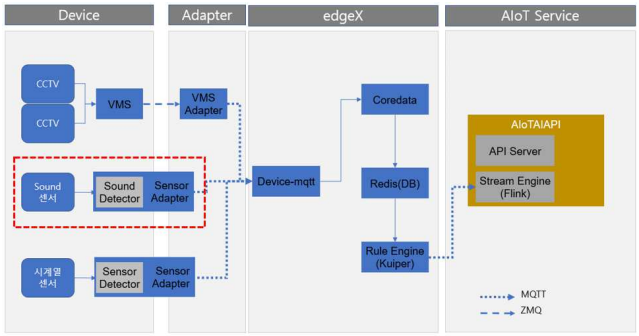

Fig. 2. device adapter architecture

B. Applied Model

This study employs YAMNet, a pretrained audio event detection model developed by Google, as the primary detection engine. YAMNet is trained on the AudioSet dataset and can classify over 500 audio categories. Within this system, YAMNet is used to detect key safety-related events such as screams, horns, sirens, dog barks, and explosions.

- Architecture: Based on the MobileNet framework, YAMNet is designed as a lightweight neural network that processes Mel-spectrogram inputs and outputs probability distributions over multiple classes.
- Input: The model takes 16 kHz mono audio signals, applies STFT and Mel filter banks internally, and produces 64-band Mel-spectrogram features.
- Output: For each frame, YAMNet produces probability scores across 521 classes, which are then aggregated to determine the final detected event.
- Edge Optimization: Due to its MobileNet architecture, YAMNet is computationally efficient and can be deployed on edge devices such as Edge TPUs, GPUs, and ARM-based processors. Depending on the device specifications, additional optimizations (e.g., quantization, knowledge distillation) may further improve inference speed without significantly compromising accuracy.

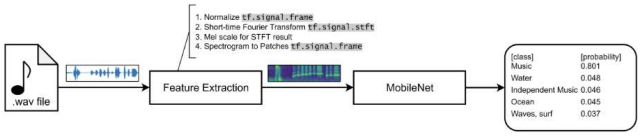

Fig. 3. A Architecture of YAMNet

C. Event Categories for Audio Context Detection

Audio event detection plays a central role in complex situational awareness for smart cities. Urban environments are acoustically rich, filled with traffic noise, crowd sounds, and machinery, which makes it difficult to reliably distinguish hazardous events from background noise. This study categorizes audio events into human-related events,

vehicle-related events, and environment- or disaster-related events, each of which is directly linked to urban safety.

- Human-related Events: Human acoustic signals provide direct cues for emergencies. Screams, shouts, crying, and groaning are strongly associated with crimes, accidents, and medical incidents. For instance, consecutive screams detected in crowded public areas may signal violent activity or accidents. These events are of high urgency and critical for initiating rapid response..
- Vehicle-related Events: Acoustic signals from vehicles are essential for traffic management. Car horns, collision sounds, tire skidding, and emergency sirens (police, ambulance, fire trucks) indicate abnormal traffic flow or accidents. Repeated horn detection in a short time window may indicate congestion or aggressive driving, while a collision sound provides a direct indicator of an accident.
- Environment- and Disaster-related Events: Rare but high-risk events such as explosions, glass breaking, gunshots, or fire alarms are crucial for urban disaster response. Immediate detection of these signals is vital for rapid evacuation, alarm systems, and disaster management operations.

This classification provides a risk-based hierarchy, enabling the system to prioritize events and map them into the complex situational awareness pipeline.

V. Performance Evaluation and Demonstration

A. Performance Evaluation

The proposed audio event-based situational awareness system aims to efficiently detect various urban safety events. In particular, the YAMNet-based audio event detector successfully identified five major hazardous sounds: scream, siren, shout, dog barking, and vehicle horn. These event types are closely associated with urban safety scenarios such as emergencies, traffic hazards, and social conflicts. Experimental results demonstrated that the system was able to reliably detect these events even under noisy urban conditions, showing the robustness of audio-based detection for real-world applications.

These findings indicate that audio data can effectively capture critical safety-related events, even when other modalities such as video may face limitations. However, complex situational awareness cannot be achieved by audio detection alone. Therefore, final validation will be carried out through integration with multimodal sensor data and user demonstration scenarios.

TABLE I. Audio Event Detection Results

Event type	Description
Scream	High-pitched human voice indicating an emergency
Siren	Emergency vehicle signals (police, ambulance, fire truck)
Shout	Loud human voice in quarrels or crowded situations
Dog Barking	Animal sound indicating potential threats
Vehicle Horn	Warning signal in traffic accidents or emergencies

B. Service Verification

To validate the practicality of the system, service verification will be conducted based on real urban safety management scenarios. Two representative cases are considered. The first is the detection of physical conflicts and fights in public areas, where the system will analyze audio cues such as shouting and abusive language, combined with behavioral patterns such as rapid movement and physical collisions, to enable early detection. The second is the detection of abnormal crowd behavior, where sudden mass movements, elevated noise levels including screams and urgent voices, and auxiliary sensor data such as temperature and smoke detection will be analyzed to identify potential fire or accident risks. Through these scenario-based evaluations, the system is expected to demonstrate its effectiveness in supporting intelligent urban safety control.

VI. Conclusion

This paper proposed a video and audio data-based situational awareness system designed for intelligent urban control, which supports the application and extension of various AI models for complex situational awareness. The proposed system was developed to efficiently detect hazardous events occurring in urban environments and to recognize complex situations based on these detections. Through this approach, the system demonstrated its potential to overcome the limitations of conventional video-centric safety management systems, such as visibility constraints and real-time processing challenges.

The contributions of this study are twofold. First, it demonstrated that applying AI models at the edge level enables real-time event stream analysis for situational awareness. Second, it highlighted the potential of combining audio and environmental data to enhance the efficiency of urban safety management. Furthermore, the proposed architecture is designed with scalability and flexibility, making it adaptable to a wide range of intelligent city safety applications.

Future work will focus on deploying the system in real urban control environments for large-scale user demonstrations and enhancing multimodal data fusion to enable more precise and robust situational awareness. In addition, by leveraging distributed intelligent edge architectures, the system will be further advanced to seamlessly integrate with diverse urban infrastructures, ultimately contributing to the evolution of next-generation smart city safety management systems.

Acknowledgment

This work is supported by the Korea Agency for Infrastructure Technology Advancement(KAIA) grant funded by the Ministry of Land, Infrastructure and Transport(Grant: RS-2022-00155803).

REFERENCES

[1] Thalluri, Lakshmi Narayana, et al. "Artificial intelligence enabled smart city IoT system using edge computing." *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE, 2021.

[2] Chiang, Mung, and Tao Zhang. "Fog and IoT: An overview of research opportunities." *IEEE Internet of Things Journal* 3.6 (2016): 854-864.

[3] Abeywardena, Yasas, et al. "Edge computing–based real-time vehicle incident detection for intelligent transportation systems." *IEEE Transactions on Intelligent Transportation Systems* 23.7 (2022): 6904-6916.

[4] Tuli, Shreshth, et al. "HealthFog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments." *Future Generation Computer Systems* 104 (2020): 187-200.

[5] Haque, Md Abu, et al. "Audio–visual surveillance: A review." *Sensors* 19.22 (2019): 4814.

[6] Xia, Xiaohui, et al. "Edge intelligence for real-time audio event detection in smart cities." *IEEE Internet of Things Journal* 8.9 (2021): 7345-7356.