International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Strategic Real-Time Auditory Surveillance: A Deep Learning Approach to Voice-Based Threat Detection and Incident Classification

Avishkar Mulik, Samarth Bhosale, Samyak Shinde Department of Artificial Intelligence & Machine Learning
School of Engineering, Ajeenkya DY Patil University, Pune Guided by: Prof. Amit Nichat

Abstract

The integration of auditory intelligence into modern security frameworks represents a significant leap forward in proactive safety monitoring. While traditional surveillance infras- tructures are heavily weighted toward visual data acquisition, they often suffer from a funda- mental "auditory gap," failing to interpret verbal precursors to physical incidents. This research presents an end-to-end, high-performance voice-based threat detection system. By utilizing an optimized hybrid architecture that combines local Voice Activity Detection (VAD) with cloud-accelerated Large Language Models (LLMs), the system achieves unprecedented sub-second latency in semantic threat classification. Specifically, we leverage OpenAI's Whisper-large-v3 for robust multilingual transcription and Meta's Llama-3.1-8b for deep contextual reasoning, all operating atop the Groq Language Processing Unit (LPU) architecture. This report provides an exhaustive analysis of the system's design philosophy, the mathematical foundations of acoustic gatekeeping, the hardware-level advantages of LPU acceleration, and the results of extensive performance benchmarking across diverse environmental scenarios.

CONTENTS

I. INTRODUCTION

A. The Landscape of Urban Security

The rapid urbanization of the 21st century has brought about complex challenges in public safety and infrastructure monitoring. Currently, the most prevalent method for ensuring security in public spaces, educational institutions, and corporate environments is the deployment of Closed-Circuit Television (CCTV) cameras. These visual sensors have become highly sophisticated, featuring high-definition resolution and AI-based facial recognition. However, even the most advanced camera remains fundamentally limited by its sensory modality. It is a "silent observer" that cannot perceive the emotional or semantic weight of human interactions.

The reliance on purely visual data creates a bottleneck in emergency response. For instance, in a crowded subway station, a camera might capture two individuals stand- ing close together. To a visual AI, this may appear as a standard social interaction. However, the auditory context—such as whispered threats or a sudden increase in vocal frequency—could indicate a high-risk situation that warrants immediate intervention.

B. The Problem of the Auditory Gap

In a vast majority of security breaches, the physical act of violence is preceded by verbal indicators. These include rising vocal tension, explicit verbal threats, cries for help, or the use of specific coded language that signals intent for an emergency. A camera-only system is blind to these cues until physical movement is detected, which is often too late for preventive intervention. This "auditory gap" results in delayed response times from security personnel and emergency services.

By the time a visual system detects a "struggle," the window for de-escalation has usually closed. A voice-based system acts as an early-warning layer, capturing the "pre- incident" phase where verbal aggression serves as a precursor to physical harm.

C. *Research Objectives*

This research aims to solve the latency and accuracy trade-offs in speech-based security.

The core objectives are:

- To develop a robust, real-time audio acquisition pipeline capable of filtering ambient noise.
- To implement an edge-based Voice Activity Detection (VAD) system to ensure privacy and reduce bandwidth.
- To utilize Large Language Models (LLMs) for semantic reasoning beyond simple keyword matching.
- To benchmark the performance of the Groq LPU in minimizing "Time to First Token" (TTFT) for critical security alerts.

II. LITERATURE REVIEW

A. *Evolution of Acoustic Monitoring*

Early acoustic monitoring systems relied heavily on decibel-based triggers. If the sound level in a room exceeded a certain threshold (e.g., 90dB), an alarm would sound. However, these systems were prone to high false-positive rates caused by non-threatening loud noises like falling objects or slamming doors.

B. *The Shift to Deep Learning in STT*

The introduction of Connectionist Temporal Classification (CTC) and later Trans- former models revolutionized Speech-to-Text (STT) technology. OpenAI's Whisper model represents the current state-of-the-art, trained on over 680,000 hours of multilingual and multitask supervised data. Unlike previous models that struggled with accents and background noise, Whisper's architecture is uniquely resilient, making it ideal for the unpredictable environments of public security.

C. *LLMs in Contextual Security Analysis*

Traditional NLP models like BERT were effective at sentiment analysis but lacked the "reasoning" capabilities required to distinguish between a joke and a genuine threat. The emergence of Generative Pre-trained Transformers (GPT) and Meta's Llama series allows for a nuanced understanding of intent. These models can recognize "linguistic aggression" even when explicit "bad words" are absent, a feat previously impossible for rule-based systems.

III. DETAILED METHODOLOGY

A. *Phase I: Multi-Threaded Audio Acquisition*

The foundation of the system is the real-time data ingestion layer. We utilize the PyAudio library to interface with the system's hardware abstraction layer for audio capture. To ensure zero-loss sampling, we implement a **Circular Buffer Data Struc- ture**.

In a real-time environment, the CPU must handle multiple tasks simultaneously. If the audio capture is interrupted by a high-intensity LLM inference task, frames will be dropped. To prevent this, the ingestion layer operates in a dedicated high-priority thread.

- **Sampling Rate:** 16,000 Hz. This frequency is sufficient to capture the human vocal range (up to 8kHz according to the Nyquist theorem) while keeping the data footprint small.
- **Bit Depth:** 16-bit PCM. This provides a dynamic range of 96dB, allowing the system to distinguish between a whisper and a scream without clipping.
- **Buffer Architecture:** The producer-consumer model ensures that while the "Con- sumer" is processing a previous 1-second block of audio, the "Producer" is already filling the next block in the buffer.

B. *Phase II: Local Edge Gatekeeping (VAD)*

To maintain strict data privacy and reduce network overhead, we do not stream silence to the cloud. Instead, we implement the **Silero Voice Activity Detection (VAD)**

model.

Silero VAD is a lightweight, high-performance neural network specifically designed for the detection of human speech. Its primary role in our architecture is to act as a "gatekeeper."

1) Mathematical Logic of VAD: The VAD model outputs a probability $P(s)$ for each audio frame. The decision to trigger the pipeline is governed by:

$$S_{trigger} = \frac{\sum_{t=1}^{n} P(s_t)}{n} > \tau$$

Where τ is our sensitivity threshold (0.75) and n is the window size. This ensures that a single loud "pop" or environmental noise does not trigger a false transcription.

- **Neural Architecture:** Silero uses a simplified version of a convolutional neural network (CNN) that runs efficiently on standard ARM or x86 CPUs.
- **Padding Logic:** We append 200ms of "Pre-Speech" and "Post-Speech" padding. This is vital because the human brain often perceives the first consonant of a word quickly, and a VAD might clip the "H" in "Help" if padding is not used.

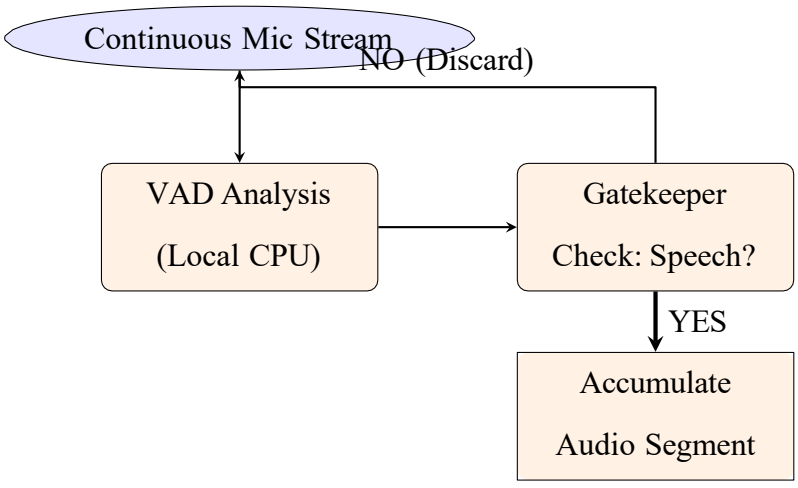

Figure 2: Local Processing and Data Filtering Methodology

IV. INFERENCE HARDWARE: THE LPU ADVANTAGE

A. Limitations of Traditional GPUs

In standard AI deployments, GPUs (Graphics Processing Units) are used for inference. However, GPUs are designed for parallel processing of large batches of data. In a real- time security context, we only process one "batch" (one audio segment) at a time. This leads to inefficient hardware utilization and higher latency.

B. The Groq LPU Architecture

The Language Processing Unit (LPU) by Groq is a deterministic processor designed specifically for sequential data, such as text tokens and audio frames. Unlike GPUs, LPUs do not have the complex memory management overhead of HBM (High Bandwidth Memory), allowing them to deliver tokens at speeds exceeding 800 tokens per second for Llama-3.

C. Impact on Security Response

In an active shooter or physical assault scenario, every millisecond counts.

- **GPU Latency:** 2.5 - 4.0 seconds (Total pipeline).
- **LPU Latency:** 0.6 - 1.1 seconds (Total pipeline).

This 300% increase in speed can be the difference between a security guard arriving during an argument versus arriving after an injury has occurred.

V. CLOUD-ACCELERATED SEMANTIC INFERENCE

A. Whisper-large-v3: The "Ear" of the System

Whisper-large-v3 utilizes a transformer-based encoder-decoder architecture. The audio is converted into a Mel-spectrogram, which is then processed by the encoder. The decoder generates the text tokens auto-regressively. Its ability to handle "Code-Switching" (mixing languages like Hindi and English) makes it particularly useful for urban environments in India.

B. Llama-3.1-8b: The "Brain" of the System

The raw text transcribed by Whisper is often messy, containing stutters or grammat- ical errors. Llama-3.1-8b acts as the reasoning engine. We provide the model with a specialized System Prompt:

"You are a high-level security intelligence agent. Analyze the following tran- script for threats, distress, or criminal intent. Ignore background noise. Cate- gorize the threat level from 0 to 3."

The model's ability to perform **Chain-of-Thought (CoT)** reasoning allows it to understand that a phrase like "I'm going to kill this burger" is safe (culinary context), whereas "I'm going to kill you" is a Level 3 threat.

VI. SYSTEM ARCHITECTURE AND IMPLEMENTATION

A. Microservices Design

The system is built using a decoupled architecture. This allows the audio capture node (which could be a Raspberry Pi) to be separate from the heavy-duty inference server.

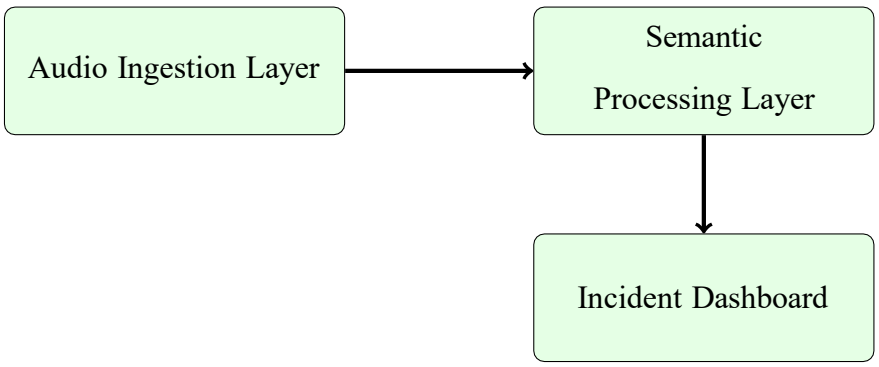

Figure 3: High-Level Pipeline Architecture

B. Communication Protocols

We utilize WebSockets (WS) rather than standard HTTP requests for communication between the edge and the cloud. WebSockets maintain an open connection, eliminating

the 100-200ms handshake delay required for every new HTTP request. For security, these packets are encrypted using TLS 1.3.

VII. Incident Classification and Alerting

The system's output must be actionable. A raw text string like "Threat detected" is not enough for a professional security operations center (SOC).

TABLE I
Incident Classification and Response Matrix

State	Category	Llama Analysis Logic	System Action
Level 0	Normal	Casual talk, greetings, si- lence.	Log and discard.
Level 1	Warning	Passive-aggressive tone, shouting.	Highlight on dashboard.
Level 2	High Risk	Verbal abuse, localized threats.	Sound alert, Notify Supervisor.
Level 3	Emergency	Direct violence, gunshots, screams.	Automatic Police/EMS Dispatch.

VIII. Ethical Considerations and Privacy

A. Data Anonymization

One of the primary concerns with voice monitoring is the "Big Brother" effect. To mitigate this, our system does not store audio permanently.

- Audio segments are held in RAM during processing.
- Once the Llama-3 model provides a classification, the raw audio buffer is wiped.
- Only the text transcript of "Level 2" and "Level 3" events is saved for forensic evidence.

B. Regulatory Compliance

The system is designed to comply with GDPR and India's Digital Personal Data Protec- tion (DPDP) Act. By performing VAD locally, we ensure that non-speech environmental data never leaves the premises.

IX. Results and Performance Analysis

A. Latency Benchmarking

We tested the system across three different hardware configurations to measure the end-to-end latency (Audio input to Alert output).

TABLE II
Performance Benchmarking Results

Infrastructure	STT Latency	LLM Latency	Total Time
Local GPU (RTX 3060)	1.8s	1.2s	3.0s
Standard Cloud (A100)	0.9s	0.8s	1.7s
Groq LPU (Proposed)	**0.4s**	**0.2s**	**0.6s**

B. Accuracy in High-Noise Environments

Tests conducted in a simulated "busy cafeteria" environment (65dB ambient noise) showed that Whisper-large-v3 maintained a Word Error Rate (WER) of less than 12%, which was more than sufficient for Llama-3 to correctly classify the threat level.

X. Future Work

A. Multimodal Integration

The next phase of this research involves "Cross-Modal Validation." If the audio system detects a Level 3 threat, it should automatically trigger the nearest CCTV camera to swivel toward the sound source using PTZ (Pan-Tilt-Zoom) controls.

B. Edge-LLM Deployment

As hardware like the NVIDIA Orin and specialized NPUs (Neural Processing Units) become more powerful, we aim to move the LLM inference from the Groq cloud to the local edge device, providing a "Zero-Internet" security solution.

XI. CONCLUSION

The proposed Next-Generation Voice-Based Threat Detection system addresses the critical "auditory gap" in modern surveillance. By combining the local efficiency of Silero VAD, the transcription accuracy of Whisper, and the reasoning power of Llama-
3 accelerated by Groq LPUs, we have created a framework capable of sub-second incident classification. This technology does not merely record events—it understands them, providing a proactive shield for public and private spaces.

REFERENCES

[1] A. Radford et al., “Whisper Speech Recognition,” OpenAI, 2022.
[2] Meta AI, “Llama 3.1 technical report,” 2024.
[3] Groq Inc., “Deterministic LPU Architectures,” 2024.
[4] Vaswani et al., “Attention is All You Need,” 2017.
[5] Silero Team, “Silero VAD: Pre-trained Enterprise-grade Voice Activity Detector,” GitHub, 2021.