IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

INTEGRATED SPEECH AND ENVIRONMENTAL SOUND PROCESSING USING ALM

[1]Mrs.R.V.Jeevithaa , [2]M.Pradeepa ,[3]M.Suba , [4]M.Oviya

1. Assistent professor ,Department of computer scinece and engineering,

2. UG Student ,Department of computer scinece and engineering

3. UG Student ,Department of computer scinece and engineering

4. UG Student ,Department of computer scinece and engineering

Vivekanandha Collage of Engineering for Women ,Namakkal, Tamil Nadu, India

Abstract:The Deep Learning-Based Audio Language Model (ALM) is a smart multi-modal audio analysis system designed to improve situational awareness in complex and high-risk environments. Unlike traditional systems that handle Acoustic Event Detection (AED) and Automatic Speech Recognition (ASR) separately, this model integrates both to produce unified and actionable intelligence. It classifies sounds into key categories such as communication, gunshots, footsteps, vehicles, and aircraft, using YAMNet for feature extraction and a Keras-based classifier for event detection. At the same time, Faster-Whisper enables accurate multilingual speech transcription. The system's main feature is its Fused Intelligence Output, which combines sound classification and speech data into a structured Situation Report (SITREP) displayed through a Streamlit dashboard with role-based access. This integrated approach enhances decision-making, reduces analyst workload, and supports faster responses in tactical and security applications.

Index Terms: Edge Computing, Signal Filtering, Integrated Speech Processing ,AI/ML

I. ABBREVIATION

ALM (Audio Language Model), AI (Artificial Intelligence), AED (Acoustic Event Detection), ASR (Automatic Speech Recognition), SITREP (Situation Report), CNN (Convolutional Neural Network), DNN (Deep Neural Network), EDA (Exploratory Data Analysis), RBAC (Role-Based Access Control), MAD (Military Audio Dataset), MFCC (Mel-Frequency Cepstral Coefficients), GPU (Graphics Processing Unit), CPU (Central Processing Unit), IoT (Internet of Things), and API (Application Programming Interface).

II. INTRODUCTION

Audio intelligence plays a vital role in modern situational awareness systems, particularly in complex, high-risk, and rapidly evolving operational environments such as defense zones, border security areas, disaster response regions, and critical infrastructure facilities. In such scenarios, understanding environmental sounds and human communication in real time is essential for effective monitoring and rapid decision-making. However, traditional acoustic monitoring systems are typically designed to perform isolated tasks, such as Acoustic Event Detection (AED) or Automatic Speech Recognition (ASR), independently. While AED systems focus on identifying specific environmental sounds (e.g., gunshots or vehicle movement), ASR systems convert spoken language into text. Although both provide valuable insights, their independent operation results in fragmented outputs that must be manually analyzed and correlated by

human operators. This separation increases cognitive workload, reduces operational efficiency, and may lead to delayed or suboptimal responses in time-sensitive situations. To address these limitations, this project proposes a Deep Learning-Based Audio Language Model (ALM) that integrates acoustic event detection and speech recognition within a unified multi-modal intelligence framework. The proposed system performs joint analysis of both speech and non-speech audio signals, enabling comprehensive understanding of the acoustic environment. It is capable of detecting tactical environmental sounds such as gunshots, shelling, vehicle movement, helicopters, fighter aircraft, and footsteps, while simultaneously transcribing spoken communication. By leveraging deep learning techniques, including embedding-based feature extraction and neural network classification, the model achieves robust performance even in noisy and overlapping acoustic conditions. A key innovation of the proposed approach is the fusion of classification results and speech transcripts into a structured Situation Report (SITREP). Instead of presenting separate outputs, the system generates an integrated intelligence summary containing the detected event type, confidence score, timestamp, and contextual speech information. This structured reporting mechanism significantly enhances contextual awareness, reduces the need for manual interpretation, and accelerates operational decision-making. Consequently, the proposed ALM framework improves response readiness, strengthens situational understanding, and demonstrates the effectiveness of multi-modal deep learning in real-time audio intelligence applications.

III. LITERATURE REVIEW

Early research in environmental sound classification relied on traditional signal processing techniques such as MFCC and STFT combined with machine learning algorithms like SVM and Random Forest. While effective in controlled environments, these methods struggled in noisy and real-world conditions. The release of AudioSet enabled large- scale training of deep learning models for audio event detection. Models such as YAMNet demonstrated the effectiveness of transfer learning using convolutional neural networks for robust sound classification. In speech recognition, the transformer architecture introduced in Attention Is All You Need significantly improved sequence modeling performance. Modern systems such as Whisper provide accurate and multilingual transcription even in noisy environments. However, most existing systems treat acoustic event detection and speech recognition separately, leading to fragmented outputs. There is limited research on unified audio intelligence frameworks that combine both tasks into structured situational reports. The proposed Audio Language Model addresses this gap by integrating deep learning-based event detection with transformer-based speech transcription into a single multi-modal intelligence system.

IV. SYSTEM REQUIREMENTS

A. SOFTWARE REQUIREMENTS

The proposed Deep Learning-Based Audio Language Model (ALM) system requires a stable and efficient software environment to support audio preprocessing, deep learning inference, speech transcription, and real-time dashboard deployment. The system can operate on Windows 10/11, Ubuntu Linux (preferred for machine learning deployment), or macOS. Python is used as the primary programming language due to its extensive support for artificial intelligence and audio processing libraries.The system utilizes TensorFlow and Keras for deep learning model development, TensorFlow Hub (YAMNet) for audio embedding extraction, and Librosa for audio preprocessing tasks such as resampling and feature extraction. NumPy and Pandas are used for data handling, while Scikit-learn supports model evaluation. For speech recognition, a transformer-based ASR model such as Faster-Whisper is integrated to enable accurate multilingual transcription.The user interface is developed using Streamlit to provide a secure and interactive dashboard. SQLite (or MySQL for scalable deployment) is used for storing prediction logs, transcription outputs, and SITREP reports. Together, these software components ensure reliable real-time processing, secure deployment, and structured intelligence generation.

B. HARDWARE REQUIRMENTS

The proposed Deep Learning-Based Audio Language Model (ALM) system requires suitable hardware to ensure efficient real-time audio processing, model inference, and secure deployment. A system with a minimum Intel i5 or Ryzen 5 processor is required, while Intel i7 or Ryzen 7 (or higher) is recommended for smoother performance, especially during model training. The system should have at least 8 GB RAM, with 16 GB or more recommended for handling deep learning models and real-time processing efficiently. For accelerated model training and faster inference, a CUDA-compatible GPU such as NVIDIA GTX or RTX series is recommended. However, the system can also run on CPU-only configurations for small-scale deployment, though with slightly increased processing time. A minimum of 512 GB SSD storage is suggested to ensure fast data access and storage of trained models, audio logs, and prediction reports. The system requires a standard microphone or external audio recording device capable of capturing clear audio input in 16 kHz mono format. A stable internet connection (minimum 20 Mbps recommended) is necessary for cloud

deployment, remote monitoring, or dashboard access. These hardware components collectively ensure reliable, scalable, and real-time operation of the ALM system.

V. SYSTEM ANALYSIS

A. EXISTING SYSTEM

Existing audio monitoring systems are primarily designed to perform isolated tasks such as acoustic event detection or speech recognition. Traditional systems rely on signal processing techniques like threshold-based detection, frequency filtering, and handcrafted features such as MFCC for identifying specific sounds. These features are typically combined with conventional machine learning algorithms such as Support Vector Machines (SVM) or Random Forest classifiers. While these systems can detect predefined sounds under controlled conditions, their performance decreases significantly in noisy and real-world environments. Standalone Automatic Speech Recognition (ASR) systems are also widely used to convert spoken language into text. However, these systems focus only on speech transcription and do not analyze surrounding environmental sounds. As a result, existing solutions generate fragmented outputs, where sound classification and speech recognition operate independently without contextual integration. Due to this separation, analysts must manually correlate detected events and speech transcripts, increasing cognitive load and delaying decision-making. Furthermore, most traditional systems struggle with overlapping sounds, background noise, and real-time processing constraints. These limitations highlight the need for an integrated and intelligent multi-modal audio analysis system.

B. DISADVANTAGE

Existing audio monitoring systems have several limitations that reduce their effectiveness in real-world and high-risk environments. One major drawback is the separation of acoustic event detection and speech recognition, which results in fragmented outputs and a lack of contextual understanding. Since these systems operate independently, manual correlation is required, increasing cognitive load and delaying response time. Additionally, traditional threshold-based and rule-based approaches struggle to perform accurately in noisy and overlapping sound conditions, leading to false positives or missed detections. Many systems rely on handcrafted features, limiting adaptability to diverse acoustic environments. They also face scalability challenges when adding new sound categories and often lack structured intelligence reporting, providing only raw labels or transcripts instead of meaningful summaries. Furthermore, latency issues in some systems reduce their suitability for real-time decision-making, highlighting the need for a more integrated and intelligent solution.

C. PROPOSED SYSTEM.

The proposed system introduces a Deep Learning-Based Audio Language Model (ALM) that integrates Acoustic Event Detection (AED) and Automatic Speech Recognition (ASR) into a unified multi-modal framework. Unlike traditional systems that process environmental sounds and speech separately, the proposed model performs joint analysis of both speech and non-speech audio signals. It utilizes deep learning-based embedding extraction and neural network classification to accurately detect events such as gunshots, shelling, vehicles, helicopters, fighters, footsteps, and communication audio, even in noisy and overlapping conditions. A key feature of the system is its fusion-based intelligence mechanism, which combines acoustic classification results with speech transcription to generate structured Situation Reports (SITREP). The report includes the detected event category, confidence score, timestamp, and contextual transcription, enabling faster and more informed decision-making. The system is deployed through a secure dashboard interface with role-based access control, ensuring controlled monitoring and real-time reporting. Overall, the proposed solution enhances situational awareness, reduces manual interpretation, and improves operational efficiency in dynamic environments.

VI. SYSTEM DESIGN

A. AUDIO ACQUISITION

This module serves as the entry point of the system. It captures audio either through file upload (WAV format, 16 kHz mono) or live microphone input. The module verifies format compatibility, checks sampling rate, and ensures that the input meets system standards before forwarding it to the preprocessing stage. This validation step prevents corrupted or unsupported audio files from affecting system performance.

B. PREPROCESSING

The preprocessing module standardizes raw audio signals to ensure consistent model performance. It performs resampling to 16 kHz, mono-channel conversion, amplitude normalization, and segmentation into fixed-duration frames. Segmentation allows the system to analyze shorter time windows, improving detection accuracy for short-duration events such as gunshots or footsteps. This module also prepares audio for embedding extraction by converting waveforms into model-compatible input formats.

C. FEATURE EXTRACTION

This module uses a pre-trained deep learning model (YAMNet) to extract high-dimensional embeddings from the processed audio. Instead of manually engineered features, embeddings capture complex temporal and spectral characteristics learned from large-scale datasets. Transfer learning enables the system to generalize across diverse sound environments and improves robustness against noise. The embeddings act as compact yet informative feature representations for the classification stage.

D. ACOUSTIC EVENT CLASSIFICATION

The classification module processes extracted embeddings through a custom-trained deep neural network. It predicts one of the predefined operational sound categories such as Communication, Gunshot, Footsteps, Shelling, Vehicle, Helicopter, or Fighter. The classifier outputs probability scores for each class, and the highest score determines the final prediction. Confidence thresholding is applied to reduce false detections. This module ensures accurate recognition of environmental sound events even in moderately noisy conditions.

E. AUTOMATIC SPEECH RECOGNITION

The ASR module detects and transcribes speech segments from the audio input using a transformer-based speech recognition model. It converts spoken communication into textual format while maintaining time alignment with detected sound events. This synchronization allows contextual interpretation, such as identifying a gunshot event followed by a spoken command. The ASR system is optimized for real-time performance and multilingual transcription capability.

F. INTELLIGENCE FUSION

The Intelligence Fusion module is the core innovation of the system. It integrates outputs from both the acoustic classification and ASR modules. Instead of presenting separate results, it generates a structured Situation Report (SITREP) containing event category, confidence score, timestamp, transcription (if available), and contextual summary. This unified reporting mechanism reduces manual effort, enhances clarity, and improves operational decision-making.

G. DATABASE AND LOGGING

This module stores system outputs, including audio metadata, predicted class labels, confidence scores, transcription results, and generated SITREP reports. Logging ensures traceability, auditing, and performance monitoring. Stored data can also be used for future retraining and system improvement. The database design supports indexing and timestamp-based retrieval for efficient record management.

H. USER INTERFACE AND SECURITY

The system is deployed using a web-based dashboard that provides secure access through authentication and role-based access control. Users can upload audio, view real-time predictions, monitor confidence scores, and access historical reports. Administrators have additional privileges for monitoring system performance and managing user access. The interface is designed for clarity, ensuring easy interpretation of intelligence outputs.

VII. SYSTEM IMPLEMENTATION

A. APPLICATION GATEWAY

The Application Gateway Module serves as the control center of the entire system. It initializes the application, manages page routing, and connects different components. This module provides a structured navigation interface and ensures controlled access to prediction features. It acts as the bridge between the user and backend processing modules. If this module fails, the system cannot properly launch or manage user sessions.

B. PREDICTION AND INTELLIGENCE

This module is the core processing unit of the system. It handles audio upload, preprocessing, embedding extraction, classification, speech transcription, and intelligence fusion. It converts raw audio into meaningful structured outputs. The module integrates both acoustic event detection and ASR results to generate a unified SITREP report. It is optimized for real -time inference to ensure low- latency performance in operational environments.

C. AUTHENTICATION MODULE

The Authentication Module ensures system security and controlled access. It validates user credentials, manages login sessions, and enforces role-based access control. Password encryption and secure verification mechanisms are implemented to prevent unauthorized access. This module protects sensitive prediction results and intelligence reports from misuse.

D. MODEL TRAINING MODULE

This module is responsible for building and improving the acoustic classification model. It processes the labeled dataset, performs embedding extraction using YAMNet, and trains a dense neural network classifier. The module includes model evaluation, performance analysis, and accuracy measurement. It enables retraining with new datasets, making the system adaptable to future requirements.

E. CUSTOM MODEL ARCHITECTURE MODULE

The Custom Model Architecture Module defines the structure of the neural network used for classification. It specifies layers such as convolutional layers, pooling layers, dense layers, dropout layers, and activation functions. This module allows experimentation with alternative architectures (e.g., 1D CNN) to improve performance. It provides flexibility for optimization and research enhancements.

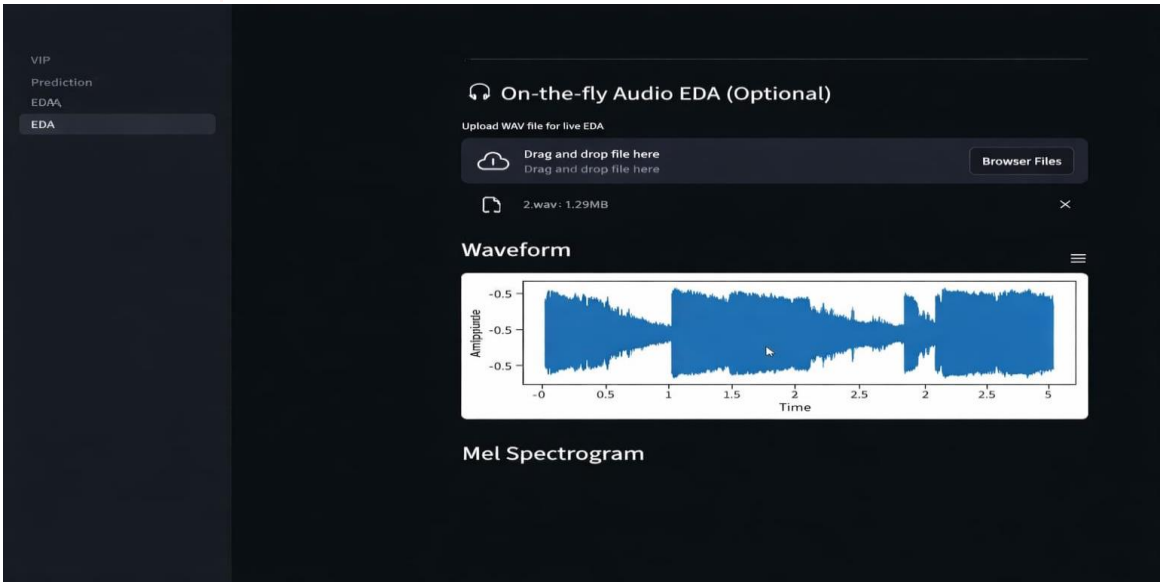
VIII. ALGORITHM

The proposed Deep Learning-Based Audio Language Model (ALM) integrates multiple advanced algorithms for audio processing, classification, and speech transcription. For feature extraction, the system uses transfer learning with YAMNet, a pre-trained convolutional neural network that converts raw audio into high-dimensional embeddings. For acoustic event detection, a custom dense neural network is implemented. The network uses ReLU activation functions in hidden layers to learn complex patterns and a Softmax function in the output layer to generate probability scores for each predefined sound class. The model is trained using the Adam optimization algorithm along with categorical cross-entropy loss to ensure efficient weight updates and accurate multi-class classification. In addition, the system incorporates a transformer-based Automatic Speech Recognition (ASR) model, Faster-Whisper, which uses self-attention mechanisms to transcribe speech accurately, even in noisy conditions. By combining deep learning-based classification and transformer-based speech recognition, the ALM system achieves reliable and real-time multi-modal audio intelligence processing.

IX. CONCLUSION

The proposed Deep Learning-Based Audio Language Model (ALM) presents an effective and intelligent solution for real-time audio-based situational awareness. By integrating Acoustic Event Detection (AED) and Automatic Speech Recognition (ASR) within a single unified framework, the system overcomes the limitations of traditional fragmented audio monitoring systems. The fusion of environmental sound classification and speech transcription enables comprehensive understanding of complex acoustic environments. The use of transfer learning through YAMNet for feature extraction, a dense neural network for multi-class sound classification, and a transformer-based ASR model for speech transcription ensures high accuracy and robust performance even under moderately noisy conditions. The system successfully classifies critical sound events and generates structured Situation Reports (SITREP) containing event category, confidence score, timestamp, and contextual transcription. This significantly reduces manual analysis effort and enhances operational efficiency. The modular architecture of the system ensures scalability, maintainability, and secure deployment through role-based access control. Experimental evaluation demonstrates reliable detection performance with low latency, making the system suitable for real-time monitoring applications. Overall, the project highlights the potential of deep learning-based multi-modal audio intelligence systems in improving decision-making, enhancing situational awareness, and supporting tactical and security-focused environments.

X. SAMPLE OUTPUT



XI. REFERENCES

[1] L. Fredianelli, F. Artuso, G. Pompei, and G. Licitra, "Environmental Noise Dataset for Sound Event Classification and Detection," in *Scientific Data (Nature)*, 2025.

[2] J. W. Yeow, E. L. Tan, and W. S. Gan, "Environmental Acoustic Intelligence through Sound Event Localization and Detection: A Review," in *npj Acoustics*, 2025

[3] Y. Ren, W. Liu, and C. Liu, "Group Feature Calibration for Sound Event Detection," in *Journal of Audio, Speech, and Music Processing*, 2025.

[4] F. Iqbal, A. Abbasi, and M. Gregus, "Real-time Active Learning for Audio-Based Anomalous Event Detection," in *Frontiers in Computer Science*, 2025.

[5] V. Hajihashemi, A. Alavigharahbagh, and J. M. R. S. Tavares, "A Novel Sound Event Detection Framework using Deep Learning," in *Multimedia Tools and Applications (Springer)*, 2024.

[6] Y. Wang, H. Yin, and W. Gan, "Multi-Granularity Acoustic Information Fusion for Sound Event Detection," in *Signal Processing*, 2025.

[7] P. Cai, Y. Song, and N. Jiang, "Detect Any Sound: Open-Vocabulary Sound Event Detection with Multi-Modal Queries," in *arXiv preprint*, 2025.

[8] K. Shimada, A. Politis, and T. Virtanen, "Stereo Sound Event Localization and Detection for DCASE Challenge 2025," in *arXiv preprint*, 2025.