ECE 445
Spring 2026

Project #80: Edge-AI based audio classifier

Team Members:
Ahaan Joishy (ajoishy2)
Kavin Manivasagam (kmani4)
Om Dhingra (omd4)

TA: Weijie Liang

1. Introduction

1.1 Problem

Most embedded audio-based systems rely on single sensors and simple threshold-based logic to detect sounds. Such approaches are highly sensitive to environmental noise and cannot distinguish between different types of sounds or identify the direction of a sound source. Cloud-based audio processing can overcome these limitations but introduces latency, privacy concerns, and increased power consumption. Therefore, there is a need for a low-power embedded system capable of performing real-time sound classification and sound source localization entirely on-device under tight memory and computational constraints.

1.2 Solution

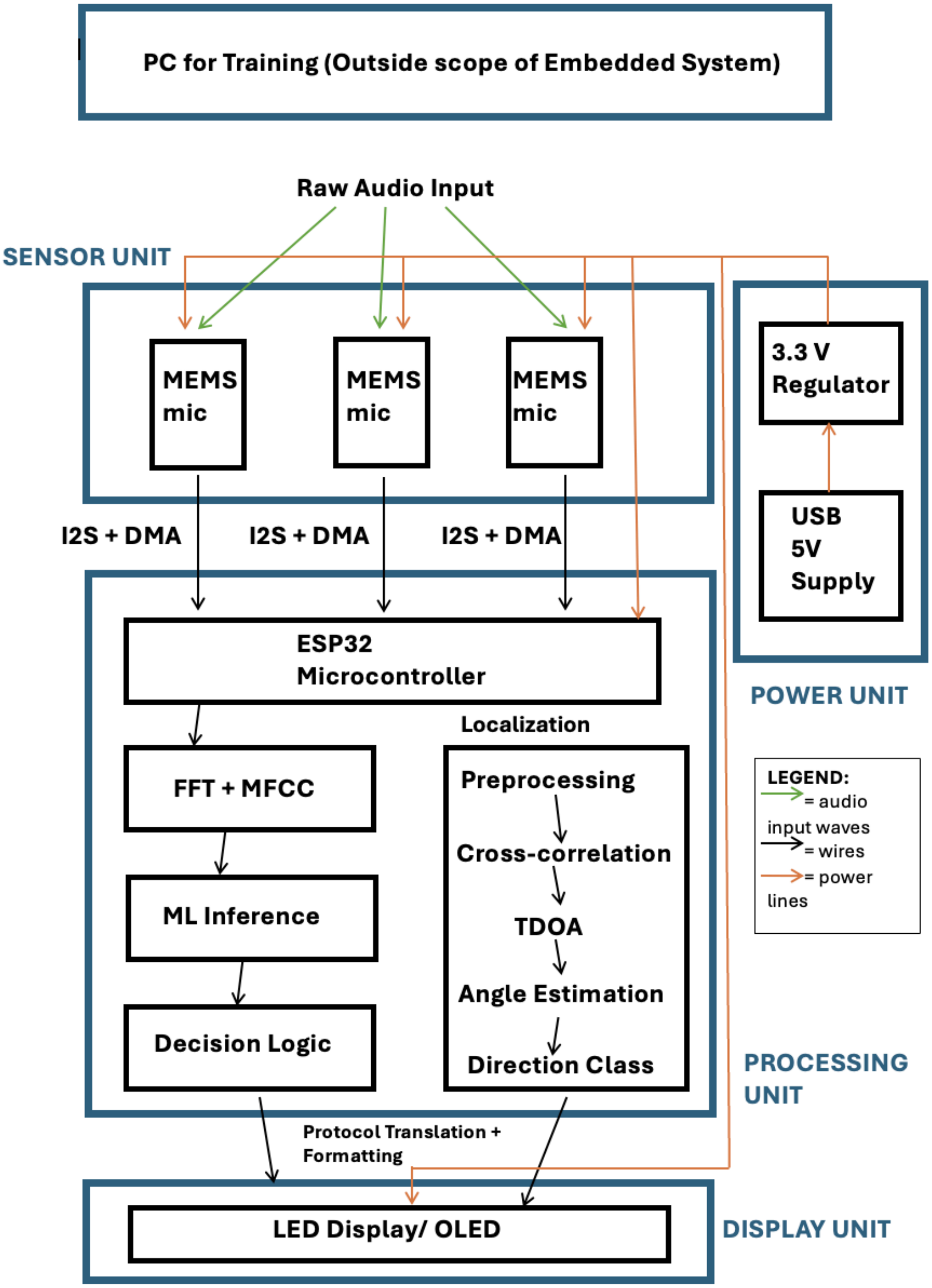
This project proposes an Edge-AI embedded system capable of performing real-time sound classification and sound source localization using multiple digital microphones and a neural network deployed on a low-power microcontroller. The system extracts spectral features from incoming audio signals, estimates the direction of arrival of the sound source, and classifies the sound into predefined categories. The results are displayed using a directional LED array or small display, allowing the user to visually identify both the type and location of the detected sound. All processing is performed locally without relying on cloud services, demonstrating the feasibility of embedded machine learning under strict power and memory constraints.

1.3 High Level Requirements

- The system shall correctly classify at least three distinct bird species vocalizations with a minimum accuracy of 85% on a recorded test dataset collected in varied environmental conditions.
- The system shall perform real-time sound localization and classification with an end-to-end latency of less than 100 milliseconds from sound detection to visual display output.
- The device shall operate entirely on-device without cloud computation while maintaining stable operation under a maximum average current consumption of 150 mA from a 5V USB power source.

2.1 Design

2.1.1 Block Diagram



2.2 Subsystem Overview

2.2.1 Subsystem 1: Sensor Subsystem:

Two or three digital MEMS microphones with I2S interfaces will be used to capture audio signals such as claps, snaps, speech, and footsteps. The microphones will be placed at known positions on the PCB to enable estimation of time differences between received signals. Audio will be sampled synchronously at 16 kHz, which is sufficient for speech and common environmental sounds. Using digital microphones eliminates the need for analog amplification and simplifies signal conditioning.

Captured audio streams will be transferred to the microcontroller using I2S with DMA to minimize CPU overhead and ensure real-time performance.

2.2.2 Subsystem 2: Processing Subsystem:

The Processing subsystem is the main control unit of the embedded system. It gets data from the sensor subsystem and is responsible for performing signal processing and the machine learning inference, after which it sends the output control signals to the display unit.

2.2.2.1) The ESP 32 Microcontroller:

- It serves as the main processor in this system and is the 'brain; coordinating all other subsystems, performs the ML inference, executes the DSP algorithms and manages the I/O with peripherals.
- Requirements: Must operate at the sufficient clock speed to perform FFT + MFCC and neural network execution within the required latency (<100ms); must handle all I/O interfaces reliably.

2.2.2.2) DSP (Happens between sensor and processing units):

- Processes audio signals to extract relevant features for ML inference. FFT transforms the time-domain signal to the frequency domain; MFCC extracts perceptually relevant features.
- Requirements: The outputs have to match the input expected by the neural net model.

2.2.2.3) ML Inference Pipeline:
The pipeline is as follows:

The extracted MFCC features are input into a compact neural network (dense or 1D convolutional) trained to recognize different sound classes. The model will be optimized for embedded deployment with a target size under 20 kB to ensure low latency and low memory usage. TensorFlow Lite Micro will be used for inference on the ESP32. If time permits, alternative embedded ML runtimes such as ExecuTorch may also be evaluated.

2.2.2.4) Localization:

- The localisation pipeline involves preprocessing (which mainly involves windowing and bandpass filtering). This is followed by cross correlation. Then, we estimate the Time Difference of arrival between sounds coming from different mics (TDOA). Finally, we calculate the angle of the direction.
- Requirement: Must calculate the direction within accuracy of +-5 degrees.

2.2.2.5) Signal processing pipeline:

1. Audio signals from each microphone are captured via I2S using DMA and stored in memory buffers.
2. The raw audio frames are windowed and transformed using an FFT to generate frequency-domain representations.
3. Mel-frequency cepstral coefficients (MFCCs) are computed from the spectral data to form compact audio features.
4. Time differences between microphone signals are estimated using cross-correlation to determine the approximate direction of arrival of the sound source.
5. The MFCC features are fed into a small neural network to classify the sound into predefined categories (e.g., clap, snap, speech).
6. The classification result and estimated direction are combined to produce the final output.

2.2.3 Subsystem 3: Display Subsystem:

A directional LED array or small OLED display will be used to present the output to the user. The display will indicate both the detected sound type and the direction of the sound source. For example, LEDs on the left or right side of the board will illuminate to represent sound direction, while different colors or patterns will represent different sound classes. This provides intuitive visual feedback and demonstrates real-time localization and classification.

2.2.4 Subsystem 4: Power Subsystem:

The system will be powered from a 5V USB input, which will be regulated down to 3.3V using a low-dropout regulator capable of supplying sufficient current for the ESP32 and peripherals. Since the ESP32 can draw large transient currents, especially during processing, adequate bulk and decoupling capacitors will be placed near the regulator and microcontroller to maintain voltage stability. The 3.3V rail will power the ESP32, two I2S MEMS microphones, and the LED or OLED display used for sound classification and localization output. To reduce electrical noise that could interfere with audio capture, each microphone will include local decoupling capacitors and optional filtering components on its supply line. The PCB will implement a star-ground layout to separate sensitive microphone grounds from high-current digital return paths. WiFi and Bluetooth will be disabled during normal operation to reduce power consumption and prevent current spikes. Overall, the power subsystem is designed to provide clean, stable power to ensure reliable real-time ML inference and accurate sound detection.

2.3 Risk Analysis

The highest-risk component of this design is the real-time audio processing and machine learning inference pipeline on the ESP32. Specifically, extracting MFCC features from dual I2S microphones while simultaneously performing sound localization and neural network classification under memory and timing constraints presents the greatest implementation difficulty. The ESP32 has limited RAM, and improper memory management or inefficient DSP implementation could cause buffer overflows, increased latency, or failed inference. Additionally, maintaining synchronized sampling between the two microphones is critical for accurate localization; timing misalignment beyond a few samples could significantly reduce directional accuracy.

Acceptable tolerances include maintaining end-to-end latency below 100 ms and keeping classification accuracy above 85% on the test dataset. Microphone synchronization error must remain within a small fraction of the sampling period (at 16 kHz, approximately 62.5 µs per sample) to preserve localization reliability. This risk directly relates to the high-level project requirements of real-time operation, multi-class bird sound identification, and accurate source localization. If timing, memory, or processing constraints are not properly managed, the system may fail to meet its accuracy and latency goals.

3. Ethics and Safety

Ethical Considerations

This project follows the IEEE/ACM Code of Ethics by prioritizing user safety, data privacy, and responsible system design. Since the system performs on-device audio classification and does not store or transmit recorded audio, privacy risks are minimized. WiFi functionality will be disabled during normal operation to ensure no unintended data transmission occurs. The system will only classify predefined environmental sounds and will not attempt speaker identification or surveillance-based tracking. Testing will involve voluntary participants producing simple sounds in controlled environments, and no personally identifiable information will be collected, so IRB approval is not required. Testing will involve recording publicly available bird sounds and outdoor environmental audio without interacting with or disturbing wildlife. No animals will be handled, manipulated, or exposed to harmful stimuli, so IACUC approval is not required.

Safety considerations

This project presents minimal safety risks because it operates at low voltage (5V USB stepped down to 3.3V) and does not involve high-voltage components or hazardous battery chemistries. Electrical safety will be addressed by using properly rated voltage regulators, current protection (such as a fuse), and adequate decoupling to prevent short circuits or overheating. All soldering and PCB assembly will follow standard lab safety procedures, including eye protection, proper ventilation, and careful handling of hot tools. The device contains no mechanical moving parts, sharp edges, or high-temperature elements that could pose a risk to users. Since the system is powered through USB and does not use volatile chemicals or large batteries, risks to end users are minimal. Overall, the project falls under low-risk laboratory electronics work and does not require a separate lab safety manual or specialized battery safety training.

References

[1] Espressif Systems, *ESP32-WROOM-32 Datasheet*, Espressif Systems, 2023.

[2] Espressif Systems, *ESP32 Technical Reference Manual*, Espressif Systems, 2023.

[3] STMicroelectronics (or manufacturer of your MEMS mic), *I2S Digital MEMS Microphone Datasheet*, 2023. *(Replace with your actual microphone part number and manufacturer.)*

[4] TensorFlow, “TensorFlow Lite for Microcontrollers,” TensorFlow.org. [Online]. Available: https://www.tensorflow.org/lite/microcontrollers

[5] IEEE, “IEEE Code of Ethics,” IEEE, 2020. [Online]. Available: https://www.ieee.org/about/corporate/governance/p7-8.html