Surround Sound: Progress Report on an Audio-Based Environment Classification System

Everett-Alan Hood
University of Washington Bothell
Bothell, WA, USA
elhood@uw.edu

Abstract

This project for Machine Learning CSS 581 is an audio-based background environment classification system based on acoustic scene recognition research. As of writing this paper, preprocessing has been fully implemented, and model training is underway on the first of three planned classifiers for background environment detection. The preprocessing pipeline includes normalization, resampling, and dataset standardization to ensure consistent audio input. The initial model is a lightweight 2D neural network for low-level audio feature learning, and its early results will guide subsequent iterations. If effective, the same framework will be applied to two additional categories, human speech and sound events, and ultimately combined into a unified audio analysis system.

1 Introduction

Audio remains an underutilized topic of research in machine learning. While computer vision is well established, auditory perception is often sidelined despite being a critical component of human experience. Most audio-based machine learning systems today prioritize medical diagnostics, speech recognition, or entertainment, leaving broader auditory awareness underexplored. This project aims to help fill that gap.

This work draws inspiration from the DCASE challenge series on detection and classification of acoustic scenes and events [2]. In these challenges, participants classify audio scenarios across predefined categories, demonstrating the feasibility of large-scale acoustic scene recognition. However, the vision of Surround Sound diverges in three ways. First, it is designed not as a single-task classifier but as a multi-layered auditory framework capable of forming a cohesive understanding of a scene across multiple forms of audio analysis. Second, the project aims for real-time interpretation rather than offline benchmarking. Third, time permitting, the long-term goal is to incorporate predictive capabilities, anticipating likely events based solely on audio context.

For this progress report, the focus is on the first pillar of the system: background environment classification. The methods, design choices, and results from this stage will directly inform future development of the remaining pillars human speech detection and sound event recognition.

2 Progress Overview

The current development phase centers exclusively on background environment classification. Future extensions will apply the same methodology to human speech and sound events. Each pillar follows three core stages.

2.1 Data Acquisition and Metadata Handling

Raw audio recordings of everyday environments are collected, organized, and tagged using JSON-based metadata. These recordings represent a variety of indoor and outdoor settings aimed at reflecting real-world auditory experiences. Metadata tools track file paths, labels, preprocessing status, and quality checks so that problematic clips can be excluded or revisited.

2.2 Preprocessing and Standardization

Each audio file undergoes normalization, resampling to 16 kHz, and reformatting to ensure uniformity. This step removes inconsistencies and prepares the data for feature extraction. Additional diagnostics record loudness ranges, duration statistics, and failure counts, which helps identify broken or atypical clips before training. While the other two pillars will require additional tailored features, there will be substantial overlap in general preprocessing across all models.

2.3 Model Architecture and Training

The initial model is intentionally lightweight, consisting of stacked 2D convolutional layers with batch normalization, ReLU activations, and max-pooling, followed by fully connected layers. These layers extract spatial–frequency patterns from mel-spectrogram inputs, capturing both harmonic structure and transient environmental cues. The training pipeline includes configurable batch sizes, learning-rate scheduling, checkpointing, and early stopping. This compact baseline provides a stable foundation for later experiments with deeper or more specialized architectures.

3 Progress to Date

Over 2,000 raw audio recordings have been processed, validated, and standardized. The first model is currently undergoing a second full training cycle after the initial attempt produced inconclusive results due to a preprocessing oversight, which has since been corrected. The project's metadata handling utilities now automatically track preprocessing failures, duration mismatches, and label consistency, providing a clearer picture of dataset quality.

CSS 581 Project, Bothell, WA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

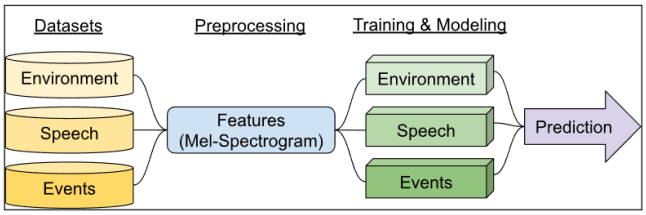

Figure 1: Overview of the Surround Sound system architecture, showing dataset inputs, preprocessing pipeline, individual models, and final prediction.

The dataset is sourced from AudioSet by Google [1], a large-scale, human-annotated collection of more than two million 10-second audio clips from YouTube. AudioSet is widely used in auditory machine learning research due to its breadth, accessibility, and category diversity. For this project, a subset focused on indoor and outdoor background environment labels was selected to align with the goal of environment classification.

Multiple mel-spectrogram configurations have been explored, varying hop length, window size, and number of mel bins to balance temporal resolution, frequency detail, and GPU memory usage. Early experiments suggest that finer frequency resolution improves discrimination between acoustically similar indoor settings, while overly large spectrograms increase training time and memory consumption.

Training is performed using Python and Jupyter notebooks, supported by libraries including PyTorch, torchaudio, librosa, NumPy, scikit-learn, jsonlines, and OS utilities. The model is trained on a 2021 Zephyrus G14 with an NVIDIA RTX 3060 Ti GPU (6 GB VRAM), a Ryzen 9 5900HS CPU, and 40 GB RAM, which together provide sufficient capacity for medium-sized experiments while imposing realistic constraints on model complexity and batch size.

4 Challenges Encountered

Challenges include dataset inconsistencies and corruption, difficulty maintaining directory structure as the project scales, and hardware limitations related to GPU memory and training speed.

Dataset Availability and Corruption. AudioSet clips originate from YouTube, so some recordings have since been removed, privated, or corrupted. Although over 5,900 clips were expected, only around 2,183 could be retrieved, and 58 of those were invalid, leaving 2,125 usable samples. This reduction required careful bookkeeping and adjustments to maintain reasonable class coverage.

Directory and File Organization. Managing the structure for three independent models quickly became unwieldy. The project folder already exceeds 3 GB, mostly due to the WAV samples stored locally. Ensuring consistency across metadata files, audio directories, and model outputs required significant reorganization and the introduction of small helper scripts to check for missing or misnamed files.

Hardware Limitations. Downloading audio from YouTube took over 12 hours, and model training has taken roughly twice that. VRAM limits on the laptop GPU restrict batch size and spectrogram resolution, slowing experimentation and forcing trade-offs between model complexity and throughput.

5 Next Steps

The immediate priority is to analyze the results of the first fully successful training cycle, including validation accuracy, loss curves, and a confusion matrix. Based on findings, hyperparameter tuning, architectural refinements, and data augmentations will be explored to strengthen model performance and robustness.

In parallel, work will begin on the next two pillars of the system: human speech detection and sound event recognition. These models will share the same preprocessing backbone while using task-specific label sets and potentially different network heads. Longer-term goals include real-time inference with live microphone input, benchmarking against DCASE-style baselines, and integrating all three models into a unified auditory intelligence system capable of providing contextual awareness in real time.

References

[1] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. 776–780 pages.

[2] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2018. Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 2 (2018), 379–393.