MINOR-2 PROJECT

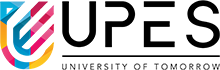
SYNOPSIS

# For

**Project Title** – Multi-Entity Detection and Counting in Audio Files using Machine Learning

**Submitted By**

|  |  |  |
| --- | --- | --- |
| **Specialization** | **SAP ID** | **Name** |
| B. Tech AIML Non Hons | 500084141 | Abhinav Singh Kushwaha |
| B. Tech AIML Non Hons | 500082410 | Aelish Kumar |
| B. Tech AIML Non Hons | 500083944 | Chirag Pawadia |
| B. Tech AIML Hons | 500083139 | Saloni Sukirti |



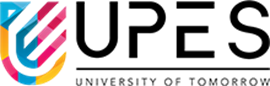
# Department of Informatics School of Computer Science

UNIVERSITY OF PETROLEUM & ENERGY STUDIES,

# DEHRADUN- 248007. Uttarakhand

Dr. Sanjeev Kumar Dr. Anil Kumar

## Project Guide Cluster Head



**School of Computer Science**

**University of Petroleum & Energy Studies, Dehradun**

## Synopsis Report

**Project Title: Multi-Entity Detection and Counting in Audio Files using Machine Learning**

**Abstract:**

In recent years, audio analysis has emerged as a pivotal area of research and application, spanning a wide range of domains, from speech recognition to environmental monitoring. This research endeavors to address a fundamental challenge within this domain: the development of an advanced machine learning model capable of detecting and quantifying multiple entities within audio files. Specifically, the entities of interest encompass dogs, cats, birds, and humans, all of which have practical significance in various fields such as wildlife monitoring, security surveillance, and acoustic scene analysis. To tackle this challenge effectively, our research adopts a multi-faceted approach, integrating signal processing techniques, feature extraction methods, and deep learning algorithms. We extract a myriad of audio features, including spectrograms, Mel-frequency cepstral coefficients (MFCCs), and chroma features. These features serve as the foundational representations of audio data for machine learning. In conclusion, our multi-entity audio analysis system represents a significant advancement in the field of audio analysis. Its potential applications span a wide spectrum, including wildlife monitoring to ensure the preservation of endangered species, security surveillance for enhanced safety measures, and acoustic scene analysis for a deeper understanding of environmental conditions. Through this research, we contribute to the growing body of knowledge and innovation in audio analysis, paving the way for more accurate and versatile applications in an increasingly audio-centric world.

Keywords - RNN(Recurrent Neural Network),Deep Learning, MFCCs(Mel-Frequency Cepstral Coefficients)

## Introduction

In recent years, the realm of audio analysis has surged in significance, finding applications in diverse fields, from speech recognition to environmental monitoring. This research project is dedicated to the development of a robust machine learning model, adept at accurately detecting and quantifying multiple entities, including dogs, cats, birds, and humans, within audio files. The approach integrates signal processing techniques, feature extraction methods, and advanced deep learning algorithms to attain high-precision entity detection and counting. This synopsis provides a glimpse into our research methodology, experimental setup, and key findings, showcasing the system's potential applications in wildlife monitoring, security, and acoustic scene analysis.

Feature extraction is pivotal in our research, where we extract various audio features such as spectrograms, Mel-frequency cepstral coefficients (MFCCs), and chroma features. These features serve as the basis for representing audio data in a format suitable for machine learning.

The core of our work revolves around the development of a multi-entity detection model, which leverages Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). This model is trained to categorize audio segments into predefined classes, such as dogs, cats, birds, and humans. Its adaptability to handle simultaneous occurrences of entities is a distinguishing feature.

In summary, this research project endeavors to create a multi-entity audio analysis system with extensive applications. From safeguarding wildlife to enhancing security measures and understanding complex acoustic environments, the system holds promise for a multitude of domains requiring automated, accurate audio analysis.

## Literature Review

Our project focuses on emotion detection from acoustic sounds. There are various research works and discoveries that support this area of study. These works have contributed to the development of algorithms and techniques for accurately detecting emotions in acoustic signals. The citations of these research papers have been reviewed and enlisted to aid our progress in the project -

* [1. ] Sruthi Kurada and Abhinav Kurada. 2020. Poster: “Vggish Embeddings Based Audio Classifiers to Improve Parkinson’s Disease Diagnosis”. The review highlights the use of VGGish embeddings for improving Parkinson's disease (PD) diagnosis through voice recordings. PD often manifests in speech abnormalities, offering potential as a diagnostic tool. Previous efforts relied on manually crafted audio features, lacking consensus and predictive power. In contrast, VGGish embeddings, known for their generalizability and efficiency, were employed. The study achieved an 87% accuracy rate for PD detection, outperforming traditional handcrafted features and competing with clinical UPDRS III-18 ratings. These findings suggest the promise of VGGish embeddings in enhancing voice-based PD classification, providing valuable insights for our audio-based entity detection and classification project.
* [2. ] Charoendee, M.,Suchato, A. and Punyabukkana, P. (2017) “Speech emotion recognition using derived features from speech segment and kernel principal component analysis,”. The review suggests that various low-level descriptors (LLDs) and features, including energy, spectral, mel-frequency cepstral coefficients (MFCC), and voicing-related LLDs, are used in speech emotion recognition. In addition, other LLDs, such as jitter, shimmer, HNR, spectral harmonicity, and psychoacoustic spectral sharpness, are also used. However, using too many features can lead to the "curse of dimensionality," resulting in lower classification accuracy. Optimal feature reduction is necessary,so we will be using mainly two of the features mentioned in the literature, namely zero crossing rate (ZCR) and mel-frequency cepstral coefficients (MFCC), for emotion detection from acoustic sounds.

## Problem Statement

In the realm of wildlife monitoring and security surveillance, the accurate detection and counting of specific entities, such as wildlife species (e.g., lions, tigers, and elephants) or human intruders, within audio recordings have become indispensable for conservation efforts and security enhancement. The problem at hand stems from the inherent complexity of audio data. In the wild, animals and humans frequently coexist, and their vocalizations can overlap, making it challenging to differentiate and count individual entities accurately. Existing systems often lack the precision required to discern between similar audio patterns, resulting in false positives or inaccurate counts.

## Objectives

Our project aims to develop a machine learning system that accurately detects and counts entities (dogs, cats, birds, and humans) within audio files. The key objectives are to achieve high precision, enable real-time processing, enhance automation, support diverse applications, and improve accuracy. Additionally, the system will be flexible enough to adapt to different scenarios and environments, making it useful for many applications.

## Methodology –

The methodology for this project will involve the following steps:

1) **Data Collection:**

* We will amass a diverse acoustic dataset encompassing human and animal sounds.
* Emotion annotations will be included to enable supervised learning for emotion detection.

2) **Feature Extraction:**

* We will extract emotion-relevant audio features, including MFCCs, ZCRs, and pitch analysis.
* These features will serve as the foundation for emotion detection.

3) **Model Training:**

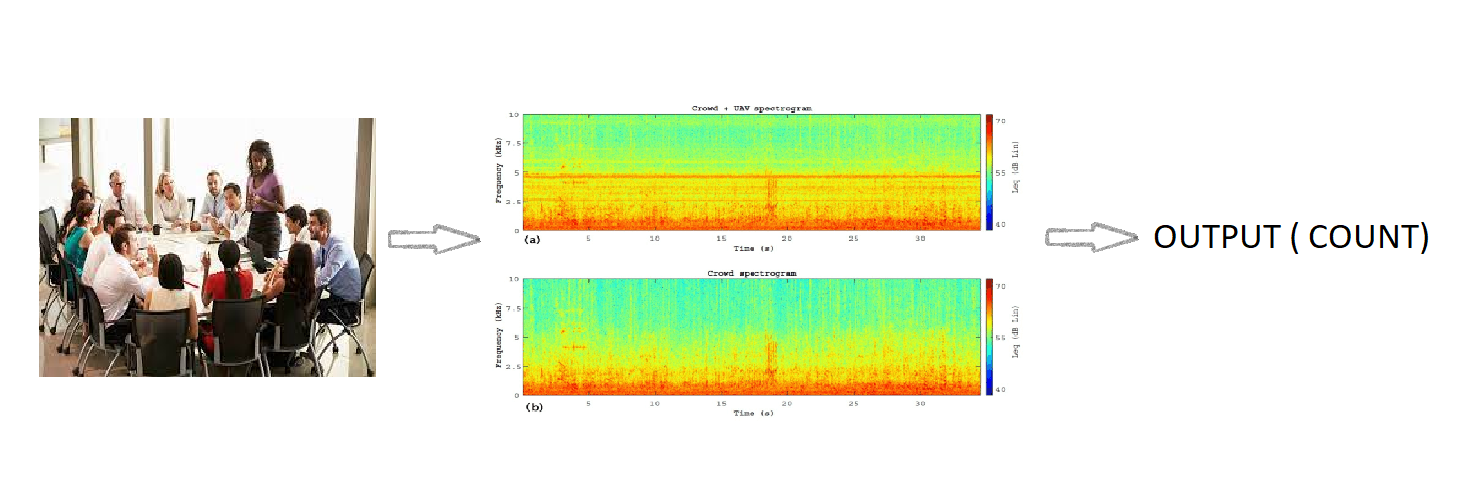
* Utilizing the annotated dataset and extracted features, we will train a machine learning model.
* The model will specialize in classifying emotions based on the extracted acoustic features.

4) **Counting Algorithm Development:**

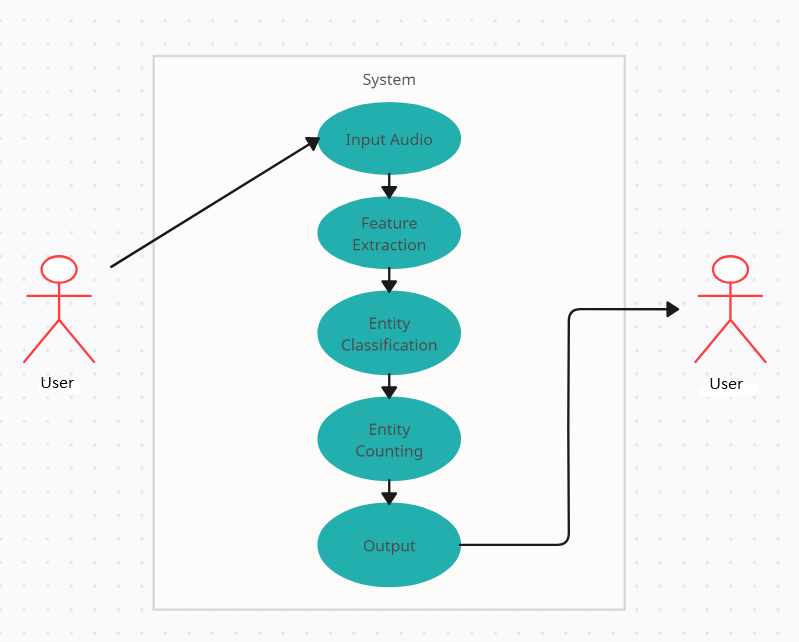
* A specialized counting algorithm will be developed and trained.
* It will accurately count detected entities, even in overlapping scenarios within audio clips.

5) **Model Validation:**

* The trained model and counting algorithm will undergo comprehensive validation.
* Evaluation will include quantitative metrics (accuracy, precision) and qualitative assessments, including result visualization.



## UML Diagram :

****

1. **PERT Chart**
2. Task 1: **Data Collection**

* Collect audio samples of different entities
* Label the data to indicate the type of entity

1. Task 2: **Feature Extraction**

* Extract relevant features from the audio data
* Clean and preprocess the data as needed

1. Task 3: **Training Classification Model**

* Choose an appropriate machine learning model for entity detection
* Train the classification model using the extracted features and labelled data

1. Task 4: **Training Counting Algorithm**

* Develop and train a specialized counting algorithm
* Integrate the counting algorithm with the classification model

1. Task 5: **Model Evaluation**

* Evaluate the performance of the combined model
* Finetune the model as needed

## References

1. Sruthi Kurada and Abhinav Kurada. 2020. Poster: “Vggish Embeddings Based Audio Classifiers to Improve Parkinson’s Disease Diagnosis
2. Charoendee, M.,Suchato, A. and Punyabukkana, P. (2017) “Speech emotion recognition using derived features from speech segment and kernel principal component analysis,”.