Contents

Contents 1

1 Introduction 2

1.1 System Introduction 2

1.2 Background of the System 2

1.3 Objectives of the System 2

1.4 Significance of the System 3

2 Overall Description 4

2.1 Product Perspective 4

2.2 Product Scope 4

2.3 Product Functionality 4

2.4 Users and Characteristics 4

2.5 Operating Environment 4

3 Specific Requirements 4

3.1 Functional Requirements 4

3.2 External Interface Requirements 8

NON Functional Requirements 9

3.3 Performance Requirements 9

3.4 Safety and Security Requirements 9

3.5 Software Quality Attributes 9

4 Design Description 10

4.1 Logical Viewpoint 10

4.2 Interaction Viewpoint 11

# Introduction

## System Introduction

Living in a world surrounded by different forms of sound from different sources, our brain and auditory system is constantly identifying each sound that it hears, in its way. Classifying audio or sound has been a major field of research for many years now and there have been many tried and tested methods with different models and features which has proven to be useful and accurate. Classification of audio can range from fields like multimedia, bioacoustics monitoring, intruder detection in wildlife areas to audio surveillance, and environmental sounds.

There are three different stages that are attached to the classification of sound; pre-processing of the audio signal, specific spectral feature extraction, and finally the classification of the audio signal. Signal pre-processing samples the input audio signal into various fragments which are utilized for extricating essential features. Zero-crossing rate, spectral flux, centroid, chroma vector, MFCC (Mel-frequency Cepstral Coefficient), and poly features are among the most well-known handcrafted features for audio classification. For the sake of our approach to audio classification, we have chosen the following features: Chroma gram, Mel spectrogram, Spectral Contrast, Tonnetz, MFCC etc. The idea of MFCC is to convert time-domain signals into frequency domain signals and use Mel filters to mimic cochlea that has more filters at low frequency and fewer filters at high frequency. Thus, it is safe to conclude that the feature MFCC and its characteristics are focused on the audibility of the human hearing system, that can accommodate the dynamic nature of true-life sounds with the way that they are treated with feature vectors for classification.

## Background of the System

An early method to handle sound recognition problem was proposed by Sawhney and Maesx, which modeled the temporal evolution of audio features and employ recurrent neural networks and a k-nearest neighbor criterion to model the mapping between features and categories. However, such temporal feature could hardly obtain an accuracy larger than 70%. Then Eronen m mn et al. discovered that the MFCC feature that based on human recognition of soundscape and how human ears discern different frequencies. Some soundscape research extract MFCC features to describe the local spectral envelop of audio signals and classify with SVM (radial basis function kernel) and random forest.

The result shows accuracy of approximately 70% which is much lower than deep learning method. Other strategy using HMM to translate an audio into events performs poorly in the presence of interfering acoustic noise. Since noise is unavoidable in practical situation, we must find features and models more robust. Therefore, researchers tend to find powerful machine learning methods like CNN, RNN with high dimensional input features such as spectrograms. These improve robustness because of the discriminative capabilities of the back-end classifiers. Our work is distinctive in the following ways. We will apply CNN with two different features: Mel and MFCC, to compare the performance of different features. And then we will try model to predict by features.

## Objectives of the System

* To identify the object in sound.
* It can be used in security systems.
* It can help in controlling traffic signals by identifying sound of traffic jam.
* It can be used in vehicles to determine if the vehicle is in running state or starting state.

## Significance of the System

This study deals with the role of sounds in object recognition in humans. Indeed, some objects are easily associated with a sound, i.e., some objects possess either a typical sound or category of sounds. This is the case, for example, of objects such as “bell” or “motorbike.” Other objects do not possess typical sounds or can be associated with particular sounds only with difﬁculty. This is the case, for example, of objects such as “table” or “pillow. ”Given that objects can be classiﬁed as a function of whether they possess or not a typical sound, a legitimate question is whether the typical sounds play any role in the visual recognition of the related objects. There are at least two opposed scenarios to frame this question.

In the ﬁrst scenario, upon the presentation of a visual object the system ﬁrst accesses an abstract representation of that object and then depending on the task at hand accesses the representations of information related to that object: among these representations is the representation of the typical sound. Thus, in this scenario, the access to the typical sound is post-categorical, in the sense that the object is ﬁrst recognized as an instance of a particular kind (e.g., a “dog”) and then the related information is retrieved. Here, the typical sound may be activated but, since its retrieval follows the identiﬁcation of the object, it does not play any role in the recognition of the object. In the second scenario, all stored representations associated to a given object are immediately and mandatorily activated upon the visual presentation of an instance of that kind of object. Here, the identiﬁcation of the object does not consist in the activation of an abstract semantic representation of this objects but instead corresponds to the activation of all stored representations. In other words, object identiﬁcation is the activation of object knowledge. So, this system is important in classification and recognition of object sounds.

# Overall Description

## Product Perspective

From the technology perspective, object sound recognition has a long history with several waves of major innovations. Most recently, the field has benefited from advances in deep learning and big data. The advances are evidenced not only by the surge of academic papers published in the field, but more importantly by the worldwide industry adoption of a variety of deep learning methods in designing and deploying speech recognition systems.

## Product Scope

The scope of this work is focused on a urban sounds classification. The elaboration of this proposal is organized as follows describe the research field of Urban. As an important processing step, feature extraction plays a critical role that will significantly affect the final classification performance. We try to break through the performance bottleneck, using novel feature sets extracted with image information retrieval techniques. It describes the experiments applying convolutional neural network (CNN), a state-of-the-art image digit recognition algorithm, to automatic extraction of sound pattern features. The system architecture, the characteristics of CNN and the classification performance are explained. studies the invariance of the widely used MFCC feature set to Urban Sounds.

## Product Functionality

* Detect Object by Voice File
* Detect Object by Recording
* Detect Object in Real-time

## Users and Characteristics

In this System user can be anything like if it is used in security system the user will the security system that will be connected to this system and generate an output by the response of this system.

## Operating Environment

This application will install on windows and minimum GPU will be required 2.17, in case of security system input device is required to accept sounds from the environment. Microphone will be used for input to the system. Or the audio file can be upload to get the result.

# Specific Requirements

## Functional Requirements

### Take voice input in Realtime

The system shall be able to take input in Realtime and shows the output on interface.

### Take voice input by audio file

The system should be able to process audio file and shows output to the interface

### Use Case Diagram

### Use Case Descriptions

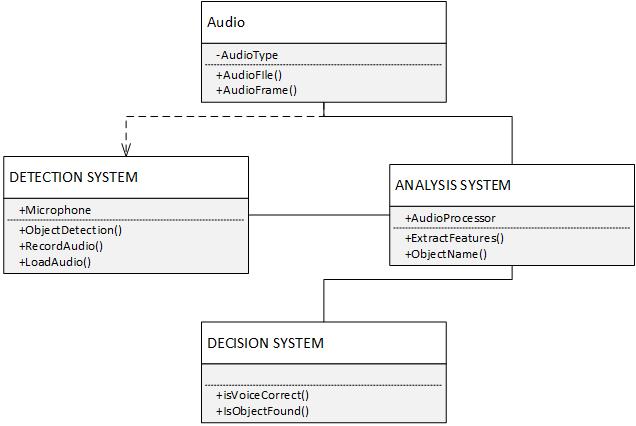
* User start or stop the software
* System will interact with the microphone while real-time listening
* System will interact with audio processing model while performing object sound analysis

### System Sequence Diagram / System Collaboration Diagram

Diagram

Description automatically generated

### Class Diagram [Optional for BSCS Students ~ Mandatory for BSSE Students]



### Data Flow Diagram



## External Interface Requirements

### Chart Description automatically generatedUser Interfaces

### Hardware Interfaces

Cutting-edge GPU is required if it will be developed in a local lab. Preferably, 5800+ Quad core GPU would be fine. 8000+ better. Since the dataset is more than 10 GB, preferably, 4 GB or more amount of RAM is needed. 1 TB + of the mass storage device is needed.

### Chart Description automatically generatedSoftware Interfaces

### Communications Interfaces

There are no external communication interfaces needed.

# NON Functional Requirements

## Performance Requirements

### Response Time

The system will work with real-life data. Response time to identifying shouldn't exceed 300 msec.

### Error Handling

When the system fails to predict in a given response time, it should give an alert message to the interface.

### Workload

The system should be able to handle identifying given input while the microphone is taking input.

## Safety and Security Requirements

There are no security requirements specified.

## Software Quality Attributes

| **Term** | **Definition** |
| --- | --- |
| Functionality | System designed to identify the objects. |
| Reliability | System should operate 99,9% in runtime. |
| Robustness | System should successfully identify urban sounds objects which will be trained with training data. |
| Portability | System should work Ubuntu 16.04 or later OS version. |
| Efficiency | Source codes of the system should be clear and operate efficiently. |

# Design Description

## Logical Viewpoint

## Interaction Viewpoint

