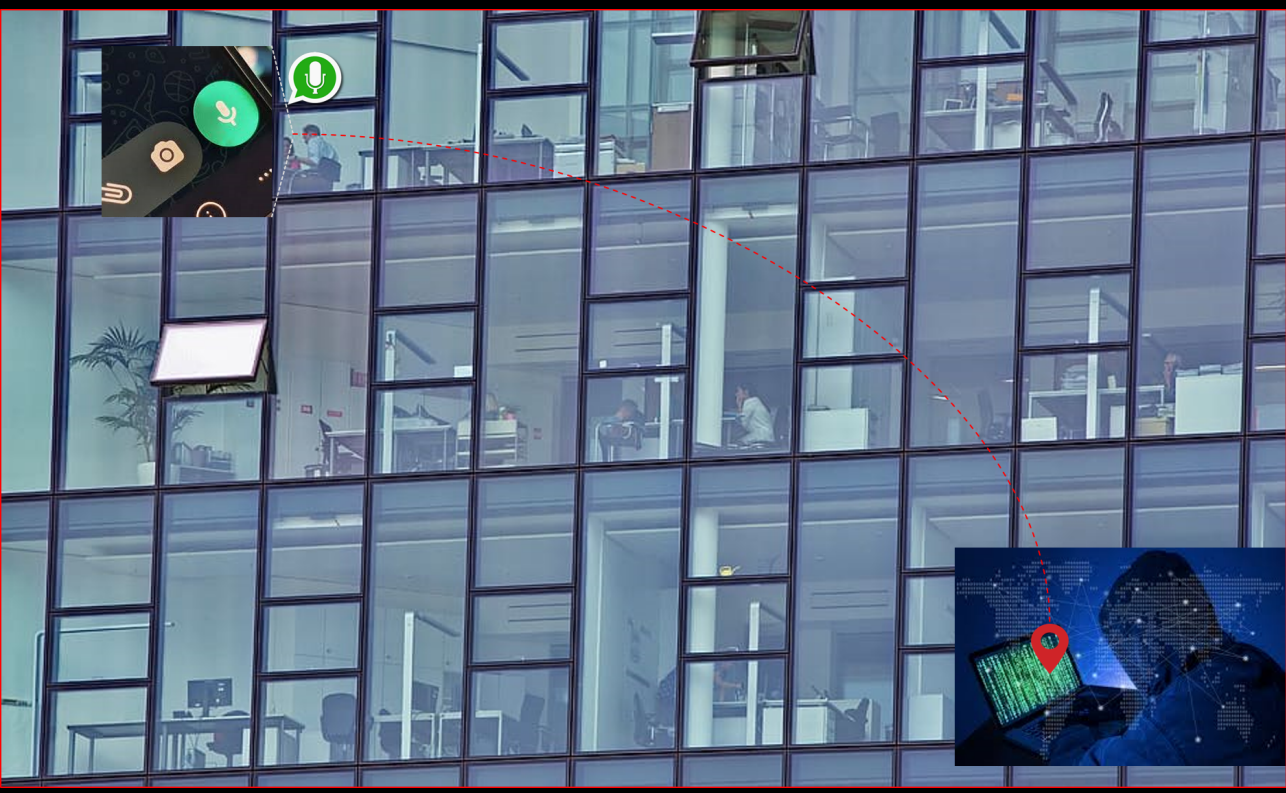
For your Voice Only

Exploiting Side Channels in Voice Messaging for Environment Detection



by

Arpita Ravindranath
Student Number: 5002702

Supervisor: Prof. Dr. Mauro Conti
Co-supervisor: Eng. Matteo Cardaioli

FOR YOUR VOICE ONLY

EXPLOITING SIDE CHANNELS IN VOICE MESSAGING FOR ENVIRONMENT DETECTION

by

Arpita Ravindranath

in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science
Specialization: Cyber Security

at Technische Universiteit Delft,
Faculty of Electrical Engineering, Mathematics and Computer Science,
to be defended publicly on August 3rd - 2021

Thesis Committee:

Prof. Dr. Mauro Conti,	TU Delft, University of Padua (Supervisor/Chair)
Prof. Dr. Odette Scharenborg,	TU Delft
Prof. Dr. Behnam Taebi,	TU Delft

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Keywords: Acoustic Side channels, environment inference, voice messaging applications

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

It is our choices, far more than our abilities that determine who we truly are.

J.K. Rowling

CONTENTS

List of Figures ix

List of Tables xiii

Preface xvii

Abstract xix

Structure of the Document xxi

1 Introduction 1

2 Related Works 5

3 Background Principles 7
3.1 Sound Waves . . . 8
3.1.1 Ambient Noise . . . 11
3.1.2 Reverberation . . . 12
3.2 Analysing Audio Waves . . . 13

4 Machine learning Classifiers 19
4.1 Audio Classifier Selection . . . 20
4.2 Classifiers Considered . . . 21
4.2.1 Linear Discriminant Analysis . . . 21
4.2.2 Logistic Regression . . . 23
4.2.3 Ridge Classifier . . . 24
4.2.4 Support Vector Machine . . . 26
4.2.5 Voting Classifier . . . 27

5 Audio Features and Audio Analysis 29
5.1 Audio Features Classification . . . 30
5.2 Feature Selection . . . 30
5.3 Audio Features . . . 31
5.3.1 Zero Crossing Rate . . . 31
5.3.2 Spectral Roll-off . . . 32
5.3.3 Spectral Flatness . . . 32
5.3.4 Spectral Centroid . . . 33
5.3.5 MFCC . . . 33

6 System Adversary Model 35
6.1 System Model . . . 36
6.2 Adversarial Model . . . 36

7 *ForYourVoiceOnly* **Attack** **39**
7.1 Data Acquisition . . . 40
7.2 Data Processing . . . 41
7.3 Model Training . . . 41
7.4 Location Inference . . . 41

8 Preliminary Experiments **43**
8.1 Different Positions . . . 44
8.2 Different Phones . . . 44
8.3 Different Audio Content . . . 46
8.3.1 Different Message Content . . . 46
8.3.2 Silent Audio Content . . . 47
8.3.3 Syllables . . . 48
8.3.4 Extracted Words from Voice Messages . . . 49
8.4 Speaker Data . . . 49
8.4.1 Phone Speaker . . . 50
8.4.2 JBL GO Speaker . . . 50

9 Experimental Settings **53**
9.1 Data Collection . . . 54
9.2 Machine Learning Models . . . 54
9.3 Cross Validation . . . 55

10 Experimental Results **63**
10.1 Location Inference . . . 64
10.2 Position Inference . . . 66
10.3 Survey . . . 68
10.3.1 Survey Organization . . . 68
10.3.2 Survey Results . . . 68

11 Conclusion **73**
10.1 Limitations . . . 73
10.2 Future Works . . . 74

A Appendix **79**

B Appendix **83**
B.1 Results for Different Positions . . . 83
B.2 Results for Different Phones . . . 84
B.3 Results for Different Audio Content . . . 85
B.3.1 Results for Syllables . . . 85
B.3.2 Results for Extracted Syllables . . . 85
B.3.3 Results for Different Message Content . . . 86
B.3.4 Results for Silent Audio Content . . . 87
B.4 Results for Speaker Data . . . 88
B.4.1 Results for Phone Speaker . . . 88
B.4.2 Results for JBL GO Speaker . . . 88

Consent Forms **89**

List of Figures

1.1 Voice propagation when sending a voice message. 2

3.1 Types of waveforms 8
3.2 Pitch of a sound 9
3.3 Difference in sonic envelope of the piano and violin on playing note C4. The image is taken from [44] 10
3.4 Sound waves of recordings taken at different positions and different locations 14
(a) *Sound wave of recording at corner P1* 14
(b) *Sound wave of recording at corner P2* 14
(c) *Sound wave of recording at corner P3* 14
(d) *Sound wave of recording at corner P5* 14
(e) *Sound wave of recording Outside* 14
3.5 Ambient noise measurements taken at different positions in a room 15
(a) *Ambient noise at corner P1* 15
(b) *Ambient noise at corner P2* 15
(c) *Ambient noise at corner P3* 15
(d) *Ambient noise at corner P4* 15
(e) *Ambient noise at corner P5* 15
3.6 Impulse Response 16
3.7 Reverberation in Speech Signal. Image borrowed from [26] 16
3.8 Example of reverberation in a room where path 1 refers to the direct sound wave and paths 2 and 3 refer to reflected sound waves 17
3.9 Segmenting a signal into frames. The image is taken from [45] 17
3.10 Sound waves of recordings taken at different positions in a room 18
(a) *Sound wave of recording at corner P1* 18
(b) *Sound wave of recording at corner P2* 18
(c) *Sound wave of recording at corner P3* 18
(d) *Sound wave of recording at corner P4* 18
(e) *Sound wave of recording at corner P5* 18

4.1 Projection of datapoints on new axis. The image is taken from [46] 22
4.2 Fitting a sigmoid function on the given data. This image is taken from [47] 23
4.3 Fitting the datpoints using Linear and Ridge Regression. This image is taken from [48] 25
4.4 Optimal line separating the classes using Support Vector classifiers. This image is taken from [49] 26

5.1 Zero Crossings in a signal 32

5.2 Spectral Rolloff . . . 32
5.3 Spectral Centroid . . . 33
5.4 MFCC Feature extraction. This image is borrowed from [7] . . . 34

6.1 Recording position . . . 36

7.1 *ForYourVoiceOnly* attack phases . . . 40

8.1 Phone position against the ear (position 1) . . . 44
8.2 Accuracy of *ForYourVoiceOnly* with different phone positions. Here task 1 is room classification between all 3 locations, task 2 is room classification between 2 indoor rooms, task 3 is position identification between all 3 locations, and task 4 corresponds to position identification between both indoor rooms. . . . 45
8.3 Accuracy of *ForYourVoiceOnly* with different audio content. Here task 1 is room classification between all 3 locations, task 2 is room classification between 2 indoor rooms, task 3 is position identification between all 3 locations, and task 4 corresponds to position identification between both indoor rooms. . . . 47
8.4 Accuracy of *ForYourVoiceOnly* with silent audio content. Here task 1 is indoor-outdoor classification and task 2 is position identification between both locations. . . . 48
8.5 Accuracy of *ForYourVoiceOnly* for task 3 when we vary the number of silent audio samples. Here task 3 corresponds to position identification within an indoor room. . . . 48

9.1 Location layout and recording positions with orientation considered in the data collection in Indoor Location I1. . . . 57
9.2 Location layout and recording positions with orientation considered in the data collection in Indoor Location I2. . . . 58
9.3 Location layout and recording positions with orientation considered in the data collection in Indoor Location I3. . . . 59
9.4 Location layout and recording positions with orientation considered in the data collection in Outdoor Location O1. . . . 60
9.5 Nested Cross Validation for Voting Classifiers . . . 61
9.6 Nested Cross Validation . . . 61

10.1 *ForYourVoiceOnly* confusion matrices for the best models. . . . 65
(a) *Complete Profiling scenario.* . . . 65
(b) *Location Profiling scenario* . . . 65
(c) *User Profiling scenario* . . . 65
10.2 Performance of machine learning models in classifying the four locations in *Complete Profiling* scenario when trained specifically with one word and all the words (i.e., combined). . . . 66
10.3 Confusion matrix for specific position inference with for I1, I2, I3 and O1 locations in *Complete Profiling* scenario. . . . 67

10.4 Gender distribution of participants of the *Complete/User Profiling* survey . 69
10.5 Age distribution of participants of the *Complete/User Profiling* survey . . . 69
10.6 Gender distribution of participants of the *Location Profiling* survey 69
10.7 Age distribution of participants of the *Location Profiling* survey 70
10.8 Comparison of task accuracy between humans and $\mathscr{F}orYourVoiceOnly$ for *Complete Profiling*. Here task 1 is indoor-outdoor classification and task 2 corresponds to the room classification between all 4 known locations . . 71
10.9 Comparison of task accuracy between humans and $\mathscr{F}orYourVoiceOnly$ for *Location Profiling*. Here task 1 is indoor-outdoor classification and task 2 corresponds to the room classification between all 4 known locations. . . 71
10.10 Comparison of task accuracy between humans and $\mathscr{F}orYourVoiceOnly$ for *User Profiling*. Here task 1 is indoor-outdoor classification and task 2 corresponds to the room classification between all 3 indoor locations. . . . 71

LIST OF TABLES

8.1 Accuracy of $\mathit{ForYourVoiceOnly}$ based on device used. Here task 1 corresponds to indoor-outdoor classification, task 2 is room classification between all three locations, and task 3 is room classification between the two bedrooms. . . . 46

10.0 Average accuracy of $\mathit{ForYourVoiceOnly}$ attack for location inference in different attack scenarios. . . . 64
10.2 Average accuracy of $\mathit{ForYourVoiceOnly}$ attack for position inference in different attack scenarios. . . . 67

A.1 The combined results of 15 participants for the Complete Profiling scenario with the AND syllable . . . 79
A.2 The combined results of 15 participants for the Complete Profiling scenario with the OF syllable . . . 80
A.3 The combined results of 15 participants for the Complete Profiling scenario with the THE syllable . . . 80
A.4 The combined results of 15 participants for the Complete Profiling scenario with all three syllables . . . 81
A.5 The combined results of 15 participants for the Location Profiling scenario with all three syllables . . . 81
A.6 The combined results of 15 participants for the User Profiling scenario with all three syllables . . . 82
A.7 Legend . . . 82

B.1 The results of Position 1 (Dataset comprising of 230 datapoints and 3 locations (2 indoor and 1 outdoor) . . . 83
B.2 The results of Position 2 (Dataset comprising of 230 datapoints and 3 locations (2 indoor and 1 outdoor)) . . . 83
B.3 The results for Different Phones (Dataset comprising of 345 datapoints and 3 locations (2 indoor and 1 outdoor)) . . . 84
B.4 The results for Same and Different Phones (Dataset comprising of 575 datapoints and 3 locations (2 indoor and 1 outdoor)) . . . 84
B.5 The results for syllable audio content (Dataset comprising of 345 datapoints and 3 locations (2 indoor and 1 outdoor)) . . . 85
B.6 The results for extracted syllable audio content (Dataset comprising of 345 datapoints and 3 locations (2 indoor and 1 outdoor)) . . . 85
B.7 Task description . . . 86
B.8 Results for different audio content (Dataset comprising of 330 datapoints and 4 locations (3 indoor and 1 outdoor)) . . . 86

B.9 Legend 87
B.10 Results for silent audio content (Dataset comprising of 265 datapoints and 2 locations (1 indoor and 1 outdoor)) 87
B.11 Results for phone speaker data (Dataset comprising of 300 datapoints and 2 indoor locations) 88
B.12 Results for JBL GO speaker data (Dataset comprising of 500 datapoints and 2 indoor locations) 88
B.13 Results for JBL GO speaker data - Testing on an unknown speaker/recordee (Dataset comprising of 200 datapoints and 2 indoor locations) 88
B.14 Results for JBL GO speaker data (Dataset comprising of 400 datapoints and 2 indoor locations) 88
B.15 Results for JBL GO speaker data - Testing on an unknown speaker/recordee (Dataset comprising of 200 datapoints and 2 indoor locations) 88

NOMENCLATURE

COCA	Corpus of Contemporary American English
I1	Indoor Bedroom I1
I2	Indoor Bedroom I2
I3	Indoor Bedroom I3
LDA	Linear Discriminant Analysis
LR	Logistic Regression
ML	Machine Learning
O1	Outdoor Location O1
OEC	Oxford English Corpus
P1	South-East corner of room
P2	South-West corner of room
P3	North-West corner of room
P4	North-East corner of room
P5	Center of room
RC	Ridge Classifier
VC	Voting Classifier

PREFACE

This thesis was written for partial fulfillment of the prerequisites of the Computer Science master's degree specializing in Cyber Security at Technical University Delft. I would like to express my sincere gratitude to my supervisors Prof. Dr. Mauro Conti and Dott. Matteo Cardaioli for their continued support, direction, criticism, advice, and redresses to the thesis work from the start till the end. I would also like to thank all of my professors, lecturers, and TA's who have taught and guided me during my master's course at TU Delft. I would like to particularly express gratitude toward Prof. Dr. Odette Scharenborg and Prof. Dr. Behnam Taebi for agreeing to be a part of the thesis committee and taking the time to study and evaluate my research work. I also want to thank every one of the 15 members who graciously took time off from their hectic schedules to be a part of my thesis project and volunteered to perform the very mundane task of recording syllables at different locations. I also appreciate the time invested by all the members of the SPRITZ research group who partook in the survey. Lastly, I would like to thank my family and friends for constantly supporting and encouraging me throughout this challenging period especially amidst a global pandemic and across different time zones.

Arpita Ravindranath
Delft, August 2021

ABSTRACT

Voice messages are an increasingly well-known method of communication, accounting for more than 200 million messages a day. Sending audio messages requires a user to invest lesser effort compared to texting while enhancing the meaning of the message by adding an emotional context (e.g., irony). Unfortunately, we suspect that voice messages might provide much more information than intended. In fact, speech audio waves are both directly recorded by the microphone, as well as propagated into the environment and possibly reflected back to the microphone. Reflected waves along with ambient noise are also recorded by the microphone and sent as part of the voice message.

In this thesis, we propose a novel attack for inferring detailed information about user location (e.g., a specific room) leveraging a simple WhatsApp voice message. We demonstrated our attack considering 7,200 voice messages from 15 different users and four environments (i.e., three bedrooms and a terrace). We considered three realistic attack scenarios depending on previous knowledge of the attacker about the victim and the environment. Our thorough experimental results demonstrate the feasibility and efficacy of our proposed attack. We can infer the location of the user among a pool of four known environments with 85% accuracy. Moreover, our approach reaches an average accuracy of 93% in discerning between two rooms of similar size and furniture (i.e., two bedrooms), and an accuracy of up to 99% in classifying indoor and outdoor environments.

STRUCTURE OF THE DOCUMENT

This document is structured as follows - Chapter 1 provides a brief introduction to the vulnerability that an attacker can target and convert to a risk. Chapter 2 discusses various works related to environment inference using audio signals and other works on location detection, these works form the basis of our research. Chapters 3, 4, and 5 overviews related background information and the necessary knowledge required by the reader to understand the $\mathscr{F}orYourVoiceOnly$ attack.The main contributions of our research start after this chapter. Next, Chapters 6 and 7 describe the system model for our attack and various attack scenarios respectively. Chapter 8 presents the various preliminary experiments conducted to finalize the setup of our experiment. Chapter 9, presents the devised setup for $\mathscr{F}orYourVoiceOnly$ attack. Then, Chapter 10 evaluates the proposed attack, discusses obtained results, the impact of audio side channels present in VoIP applications, and also exhibits practical implications of $\mathscr{F}orYourVoiceOnly$ attack. Finally, Chapter 11 summarizes the work and its limitations and proposes potential future research directions.

1

INTRODUCTION

In this chapter, we introduce WhatsApp: one of the most used messenger applications worldwide. We then discuss potential sources of information leakage present when using the voice messaging service of this application. Finally, we introduce our proposed attack and the contributions that we propose through our work.

1

Ever since its inception in 1992, the smartphone is a gift that keeps giving. It allows its users to communicate with anyone around the world on the go. With the ascent in the ubiquity of the smartphone, it makes for a lucrative target for attackers. Instant messaging applications are one of the most used apps on the phone. Modern chats have replaced feature-poor SMS by adding text, images, video, audio, and emoticons to the text. This has allowed instant messaging apps to attract more and more users over the years. In 2020, more than 2.7 billion users used at least one instant messaging app [1]. Nowadays, the most used instant messaging app with over 2 billion users worldwide is WhatsApp [2]. One of the most used features by WhatsApp users are voice messages, so much so that over 200 million are sent every day [3]. Sending a voice message requires even less effort for a user compared to texting. Moreover, voice messages allow enhancing the meaning of the message by adding an emotional context (e.g., irony). Given the appreciation of users, this feature has become common in other messaging apps as well [23], but does a voice message send more than we intend to?

As can be seen in Figure 1.1 when a person speaks the voice signals travel in different paths some of which undergo reflection. The reflected paths depend on the shape, dimension, furniture, etc. that are present in the room, and along their path, these waves are affected by phenomena such as diffraction, refraction, reflection, and interference. Reflected audio waves end up back at the speaker causing the persistence of noise which is termed as reverberation. In addition, other ambient noises are also present - such as noises from secondary audio sources. The combination of both alongside the audio message gets picked up by the smartphone during voice messaging.

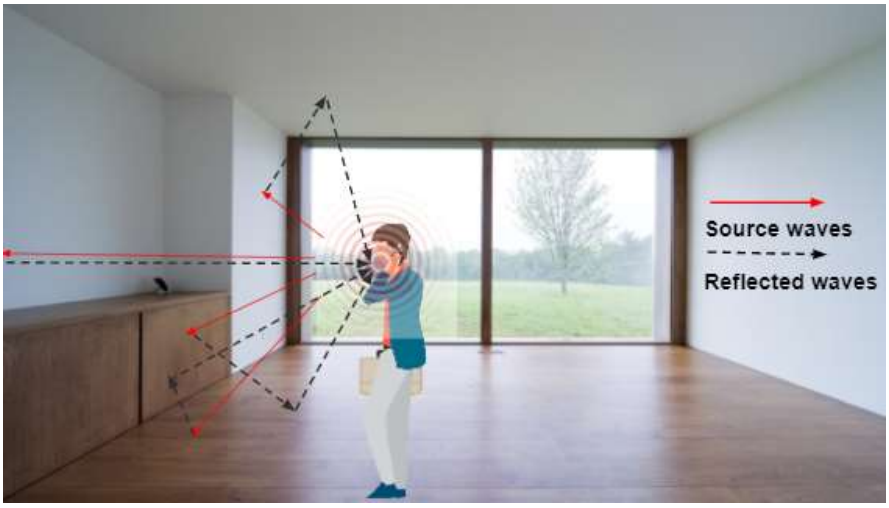


Figure 1.1: Voice propagation when sending a voice message.

[1] https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/
[2] https://www.whatsapp.com/
[3] https://www.thesun.co.uk/tech/6815812/texts-voice-messages-whatsapp-imessage-switching/

Side-channel attacks and in specific acoustic side-channel attacks are not a new genre of cyber-attacks. Historically, there has been a lot of research on such types of attacks for instance side-channel attacks on keyboards utilizing the sound produced due to various keystrokes[29], attacks on printers based on their acoustic emissions[14], attacks on smartphones due to the acoustic emissions emitted when we interact with phone screens [37] etc. We present a means to use such acoustic side channels as the premise of a new attack. We aim to make use of the physical measures mentioned that are readily accessible and inadvertently shared during WhatsApp audio messaging to gain intelligence about the victim's whereabouts. This information can be utilized in a two-sided manner based on the motivation. Both law enforcement agencies (for the purpose of forensic investigation) and attackers can benefit from the leaked information. The main contributions we propose in this research work are:

- We propose a novel attack for inferring a specific user location (e.g., a specific room) leveraging simple WhatsApp voice messages.
- We collected a dataset of 15 people and 4 different environments (i.e., three indoor one outside) for a total of 7200 recordings (i.e., 480 per participant). We will make the dataset public, available to the research community upon acceptance. We believe it will be useful in studying the problem further and developing countermeasures.
- We performed an analysis of our attack simulating three different real attack scenarios based on the knowledge available to the attacker. We demonstrated that our attack can distinguish the location of the message among a pool of known environments (i.e., three bedrooms, and a terrace) with an accuracy of up to 85%. Moreover, we showed that our approach reaches an average accuracy of 93% in discerning the voice message location of two rooms of similar size and furniture (i.e., two bedrooms). We further inferred the room and the specific position of the user within the room (e.g., a corner) for this task we achieve an accuracy of 64%.
- We conducted a study to assess human ability in discerning the location of an audio recording. We divided the survey into two parts to closely recreate the attack scenarios simulated by our model. We show that humans mainly rely on guessing and perform very poorly in comparison with our $\mathit{ForYourVoiceOnly}$ attack. Our participants were able to reach an average accuracy of only 24% in distinguishing the location of the message among the 4 known environments (i.e., three bedrooms, and a terrace).

The novel contribution of our work is the use of audio recordings (audio messages) which are compressed files with reduced file sizes to perform location inference. This technique requires no physical access and so it increases the feasibility in the real world of the attack scenarios mentioned in Section 6.2.

2

Related Works

In this chapter we provide a brief overview of the existing works related to obtaining environment information from audio. These works form the basis for the ℱ or Your Voice Only attack we proposed.

2

Sound classification represents a field of increasing interest in several areas and applications such as, surveillance [30], medicine [39], emotion recognition [40], music genre classification [31], and forensics [35]. The three main disciplines involved in sound classification are: Music Information Retrieval (MIR), [38],[42], Automatic Speech Recognition (ASR) [32, 43], and Environmental Audio Scene Recognition (EASR) [33, 41]. Music and speech can be well described by features such as MFCC (Mel-frequency cepstral coefficients), bandwidth, zero-crossing rate (ZCR), and spectral flux [11, 12]. While for the recognition of environments the problem is more challenging since the sound, in this case, does not present any tonal or harmonic structure [19].

The EASR problem involves the identification of the environment of recorded audio. A first comprehensive study on EASR was carried out by Cowling et al. [9]. In this work, the authors explore different feature extraction and classification techniques on EASR, achieving a 70% accuracy leveraging dynamic time warping classification techniques. One of the primary tasks in the EASR domain is the distinction between indoor and outdoor environments. Khonglah et al.[28] proposed the use of foreground speech segmentation to obtain foreground and background segments of an audio recording. Then from the obtained segments the MFCCs were extracted and used to train an SVM classifier to perform indoor-outdoor classification. In this study, the authors highlighted that the major cause of misclassification was the presence of speech in the background. Not only speech but also other background noises can induce classification errors. In real-world scenarios, it is quite common to have complex environment sound (i.e., environments with multiple sound sources). To mitigate the impact of complex sounds on environmental prediction performance, Delgado et al. [20] introduced a feature reduction strategy using a Chi-Squared Filter [3]. Unfortunately, a similar approach cannot be applied to the classification of similar locations. Both speech reverberation and background noise are important sources of information that can be descriptive of the environment in which the voice message is recorded.

Recently, many works on EASR have leveraged deep learning algorithms to perform feature extraction and classification [22, 25, 27, 36] Based on the work conducted by Chandrakala et al. [33] deep learning approaches show better performance compared to traditional machine learning techniques. However, these approaches cannot be applied in our case, since they require large amounts of data to train the models.

Additional factors that affects EASR are the quality of the recording device and the format in which the sound signal is saved (i.e., lossy audio formats). In this regard, several works have focused on the recognition of environments from sounds recorded with resource constrained devices (e.g., smartphones). Gomes et al. [24] presents an application for the smartphone device to classify an audio recorded on the device using a combination of SAX-based multiresolution motif discovery in combination with MFCC.The work by Peltonen et al. [8] aims to perform context-based audio scene recognition. However the data used in this work was obtained using a stereo setup and stored in a digital audio tape recorder.

There are works performing indoor location identification that make use of other data such as the GSM/Wifi signal in combination with audio recorded on a phone [18]. To the best of our knowledge, there are no works in the literature that attempts to identify a specific location (e.g., a specific room) from a voice message recorded by a smartphone.

3

BACKGROUND PRINCIPLES

In this chapter, we introduce some basic concepts about sound waves and their behavior. In the first part of this chapter, we discuss some of the fundamental characteristics of sound waves, while in the latter part, we present the behavior of sound waves when interacting with obstacles that we wish to exploit in our $\mathcal{F}orYourVoiceOnly$ *attack.*

3.1. SOUND WAVES

When a person speaks sound waves are generated due to the displacement of particles in the medium causing vibrations. These sound waves are mechanical waves that travel through the medium of transmission. These waves comprise compression (vibrating particles are closest together) and rarefaction (vibrating particles are farthest apart). Although these waves are longitudinal they are represented as transverse waves for easier representation. Sound waves can be broken down into 4 major types of contributory waveforms 3.1 - the sine wave, square wave, triangle wave, and sawtooth wave. Of these

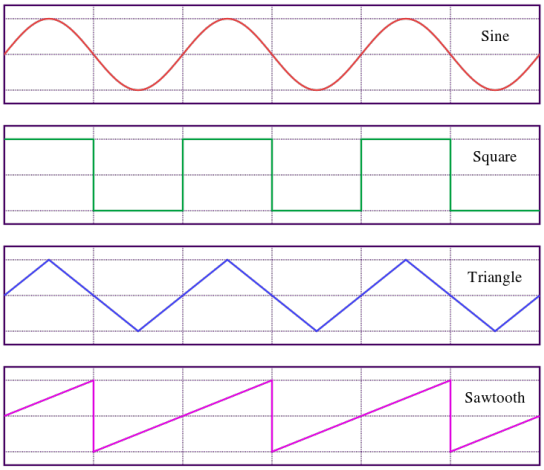


Figure 3.1: Types of waveforms

waveforms, the sine wave contains a single fundamental frequency and no overtones or harmonics. Both the square and triangular waves contain odd harmonics in addition to the fundamental sound. The sawtooth wave is the richest of the 4 waveforms in terms of timbre. Through air sound waves travel at a speed of 3.44 km per second. When these longitudinal waves reach us humans the wave is broken down into the basic elements of time and pressure. This then is used by our auditory system to process every sound we hear. Similar to any wave the sound wave also has the characteristics of frequency and amplitude attached to it. However, the analysis of sound involves some other characteristics such as:

- ***Pitch:*** The pitch of a sound or the shrillness is related to its frequency. A higher frequency is associated with a higher pitch and a lower frequency with a lower pitch. The pitch helps us order the sounds on a frequency scale and characterizes how high or low a sound is for the human ear. This makes pitch and frequency one and the same feature but they are described separately based on whether we are discussing how we perceive the sound or whether we wish to mathematically calculate a value as a physical feature. The Figure 3.2 depicts the wave for two sounds with differing pitches.

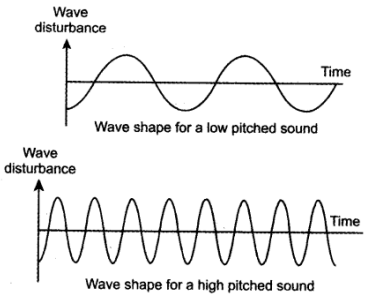


Figure 3.2: Pitch of a sound

While this depicts the pitch of a simple sound complex sounds have multiple frequencies associated with them and the perceived pitch is not the same for every listener. In general, if the constituent frequencies of the complex sound differ by about 7 Hz (this value varies per person) or more humans can detect the complex pattern obtained due to either superposition or interference of the constituent waves of different frequencies.

- ***Loudness:*** As the name suggested it's how "soft" or "loud" a sound is perceived. It is directly dependant on the amount of energy or intensity of a sound wave. This characteristic helps us to order sounds from quiet/soft to loud noises. The loudness is proportional to the square of the amplitude of the signal - hence, the more the amplitude the louder is the sound. In terms of the human auditory system, the loudness is based on the number of nerve stimulations. It is seen that higher intensity waves push the basilar membrane more resulting in more nerve stimulations[1]. This also means that a complex sound will sound *louder* than a simple sound of equivalent amplitude because more nerve stimulations occur in the case of complex sounds.

- ***Duration:*** This refers to how long a sound lasts or its time duration. This is not equivalent to how long the sound actually lasts or the "physical duration" as the sound may still persist. So it alludes to the time from when the sound is heard first till it stops or it changes. When a new wave pattern is recorded then humans perceive it as a new sound and this sound *stops* only when we find this pattern is no longer repeating. Often when multiple sounds are playing simultaneously in noisy environments it is hard to distinguish sound individually. Hence we hear many sound sources together and we perceive it to be a continuous sound while in fact in reality it is not.

- ***Timbre:*** Timbre (tone quality or tone color) is known to provide insight on the *quality* of the sound. This feature is what allows humans to distinguish between sounds and musical instruments. Timbre is a feature that reports the cumulative effect of multiple factors such as spectral envelope, noise, frequency/amplitude modulation, etc. It is also indicative of how a sound changes with time. Even if the sounds are the same in terms of frequency and loudness they can be distin-

guished due to this feature. Music pieces played on two different instruments can be distinguished due to differences such as overtones (frequencies of a waveform that are higher than, but not directly related to the fundamental frequency), sonic envelopes. These differences are reflected in the timbre. Figure 3.3 shows how different the sonic envelopes are for two different instruments (piano and violin) playing the exact same note - C4. The sonic envelope consists of the following four phases:

- *Attack (Referred to as **A** in the figure):*This phase is found right at the beginning when a key is pressed or a sound is generated till the sound reaches its peak.
- *Decay (Referred to as **D** in the figure):*During this phase the sound signal drops from its peak amplitude to a more stable sustain level.
- *Sustain (Referred to as **S** in the figure):*In this phase the sound is maintained at the sustain level till the release of the key or the sound is stopped.
- *Release (Referred to as **R** in the figure):*This phase refers to the last phase when the sound drops from sustain level to 0 or silence.

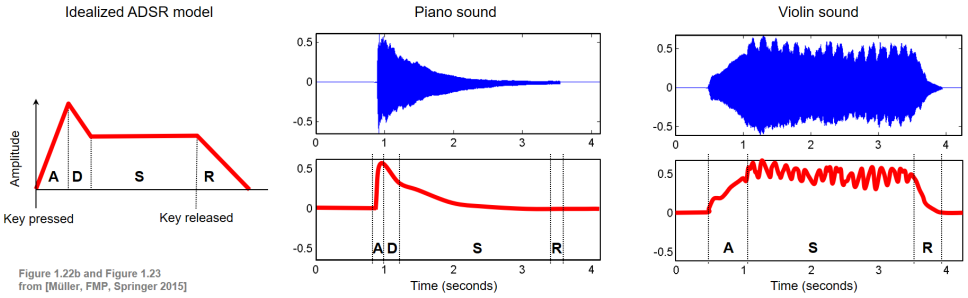


Figure 3.3: Difference in sonic envelope of the piano and violin on playing note C4. The image is taken from [44]

- ***Sonic texture:*** This feature aids in coherent distinction of sound sequences that are perceived to be produced by the same source. The *texture* is embedded within the sound and is influenced by the mental picture that the sound creates. The texture is dependant on the density, pitch range, number of sound sources, etc. For instance, the sound wave which is in the form of a sine wave stimulates our auditory system more and is termed *rich* but it has a very low harmonic content compared to say, not so smooth square or triangular waves which are symmetric about the origin. In a musical context, it refers to the quality of the melody based on how the harmonic components are combined. In music, textures can be categorized into 5 different categories monophonic, homophonic, polyphonic, homorhythmic, and heterophonic. The texture in comparison to timbre must contain some sort of dissociability in the various measurements — time, frequency, or intensity. In exceptional cases in which we can no longer separate the synchronous events into their components, the texture becomes a timbre.

- ***Spatial location:*** This feature enables us to detect the source of sound in terms of distance and direction. This characteristic helps us to build a map of the environment and place the sound source in it w.r.t the vertical and the horizontal plane. This feature helps in sound source distinction when there are multiple sources - for instance, we can identify sound from a single speaker in a noisy environment like a restaurant or pub.

As a human when we hear any sound within the hearing range (20 Hz - 20,000 Hz) we are able to process where the sound source is. The sound waves with a frequency below the hearing range are infrasonic sounds and the sounds with a frequency higher than this range are termed ultrasound. Sound waves have acoustic properties that provide the listener with cues to infer the source of the sound. These properties get distorted due to the reflections of these sound waves from the different surfaces present in the room. However, this does not affect the ability to find the source of a sound as only the direct sound that arrives first is considered for this task by the human auditory system.

Noise is commonly used in science to refer to any component that interferes and obfuscates our signal of interest. Very loosely in acoustics, we can say any sound wave which is not of interest to us is "noise". These noises can be due to various reasons such as bio-acoustic noise, noise due to the environment, noise from electronic devices, etc. In the following sections, we discuss two types of noise and their relevance in source location identification.

3.1.1. AMBIENT NOISE

Ambient noise refers to any sound which is not of primary interest and is a background sound generated by various sources. These sounds can be generated by other people,bio-acoustic noises generated by animals, noise from electric devices like the fridge, printer, heater, motors, etc, noise from the environment like rain, traffic noises, etc. These noises may also be very typical of a location. For instance, the background noises within a room are different from that outside. Some of the differences can be due to the presence of more noisy background elements outdoors such as traffic, kids playing, animal noises, wind, etc. Within a room, these noises are highly reduced and background noises may be mainly due to electronic devices, other people in the room, etc. For instance, if the indoor room is a restaurant some prominent background noises can be talking, cutlery noise, music. Another indoor location can be an office cabin where the main sources of background noises can be from the computer, the air conditioning or heating system, voices of the people usually present in the cabin. So we see that different locations may have a different combination of noises that are characteristic to it. The Figure 3.4 shows an audio recording with no speech content at different indoor positions in a room (studio room) and an outdoor location (balcony). There is some difference in the waveform of the outdoor recording in comparison with the indoor recordings. However, we cannot find a very distinguishing characteristic among the waveform of the recordings of different positions within the indoor room. Although there is no clear visual distinction possibly due to the background noise contributors being the same at the different positions within a room the loudness of these noises would still vary at the different positions which would help identify a singular position within a room. To confirm this we measured the ambient noise at various positions within a room using applications such as

sound meter, infrasound detector, and ultrasound detector. The variations in the values of the measurements at the different positions are shown in Figure 3.5.

3.1.2. REVERBERATION

The human auditory system can recognize the 3-dimensional position of a sound source - given by the distance, horizontal angle, and vertical angle. This is done with the help of direct sound obtained directly from the sound source by the listener's ears. Figure 3.8 shows different paths taken by the audio signal where path 1 refers to the direct sound wave reaching the listener which takes the shortest time as it travels the shortest distance. Along with these direct waves, there are also waves that have been reflected off the different surfaces of the room termed as *reverberations*. The human voice falls within the normal range of 500Hz to 2KHz, these sound waves generated may be reflected off even small objects if the frequency is sufficiently high. Reflections result in echoes, reverb, and standing waves. Since most locations are not soundproofed or acoustically treated many a time the reflected waves can also cause a lot of unwanted phenomena such as slap echo, standing wave, and comb filtering. Reverberations are affected by the room size and shape, room layout, furniture, materials used for construction and decor, people present in the room, etc. This is due to their position affecting the reflected waves and also the fact that these objects have varying levels of absorption coefficients varying between 0 and 1.This coefficient is representative of the number of sound waves which are absorbed vs. the amount that is reflected back in the room. An absorption coefficient of 1 means that almost no waves are reflected back and e.g. of such an object is an opening like an open window whereas a coefficient of 0 means most of the incident sound waves are reflected back and almost 0 percentage is absorbed by the object, an example of such an object is a concrete wall. When we record with a microphone very close to a huge obstruction the amplitude of the sound wave may roughly double due to the combination of the incident and reflected waves which tend to be in phase (so they constructively interfere). However, this need not be always the case and the waves interfere constructively or destructively based on their phases.

Figure 3.6 depicts the response of an impulse in a room that is not acoustically treated. The first couple of reflections are distinct echoes and correspond to the reflected waves which travelled along shorter paths. The later reflections present in the tail are indistinct in nature and correspond to the reflections travelling longer routes. While this is the behaviour of an impulse the speech wave undergoes slightly different changes in its signal with time. The direct speech signal is termed dry speech. This dry speech signal interacts with the echoes generated to form the reverberant signal as shown in Figure 3.8.

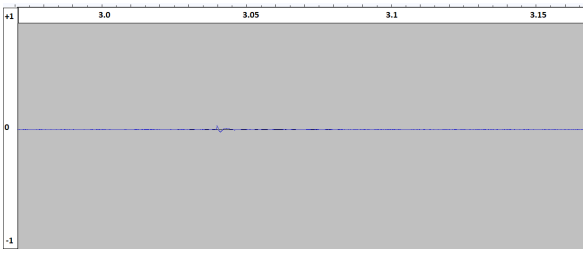
These reflected waves arriving at the listener lead to the persistence of sound even after the source has stopped generating the sound. This time it takes to for the sound to actually stop after the sound source stops is the *reverberation time*. One of the commonly used metrics to measure reverberation time is *T60* or the *reverberation time 60 dB* which corresponds to the time taken for the sound pressure level to drop by 60 dB after the source sound stops. The reflected waves are very indicative of the speaker location or source location. Path 2 and 3 in Figure 3.8 refer to reverberations wherein the path followed by the third reflection is longer than that followed by the second reflected

signal due to the reflections occurring along the path. If the gap between the time the sound starts till the first strong reflection of the sound arrives is more, we know that the reflection had to travel a longer path, which means there was no physical obstruction present close by. Another characteristic that is used even by the human auditory system is the ratio of direct sound to reflected sound. We see that these reverberations give a lot of information based on the path the reflected waves travel. Since these paths will be different for different locations and even different positions within the room - we propose to leverage this distinction in our attack. Figure 3.10 shows how the waveform of the same speech recording at different positions within a single indoor location (studio room) looks different.

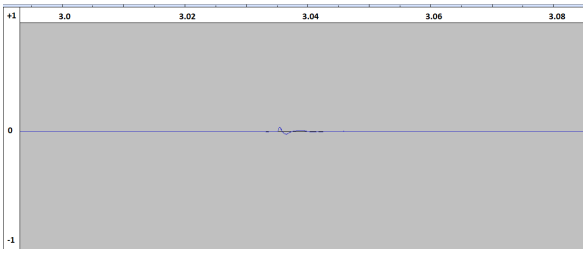
3.2. ANALYSING AUDIO WAVES

Audio is usually stored in files of formats such as Wav, Ogg, etc. Audio signals are continuous waveforms and the first step to understand these signals digitally is to convert these continuous waveforms to discrete waveforms by performing *sampling*. During discretization samples are taken from the continuous signal to calculate the discrete value - the number of samples taken per second is known as the sampling rate and sample refers to the value of the signal at a point in time/space. The higher the sampling rate the more information is retained during sampling but it is also more expensive computationally.

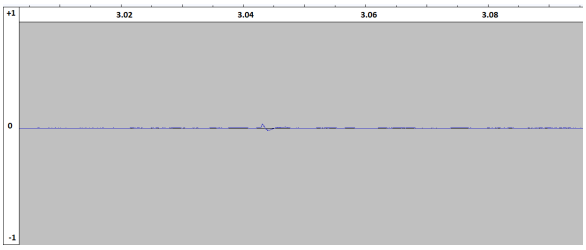
During audio signal analysis, mostly the frequency-related or spectral characteristics are considered as the amplitude variation with time is not very informative. This process requires a Fourier transform to convert the time-based signal to the frequency domain spectrum. But this results in a complete loss of time information of the non-stationary wave in favor of frequency information. To sort this issue *spectograms* are used, they represent the variations of the frequency of the signal with time - this is known as short-term analysis technology. To do so the signal needs to be broken down into smaller fragments called frames (as shown in Figure 3.9) and the Fourier transform is calculated for each of these frames. Figure 3.9 shows how a signal is segmented into frames of length N and having an overlap of length M.One of the most commonly used techniques is the STFT or the Short-time Fourier transform which performs the fast Fourier transform on the individual frames after segmenting the signal. To do so the signal that is being fragmented is multiplied by a window function - this function has a non-zero value for a very short period, so it helps retrieve a very small section of the signal. Then this window function is shifted by a distance. The presence of overlap as seen in Figure 3.9 is important so that no important occurrence at the end of the frame is lost. The features are then extracted on a frame-by-frame basis.



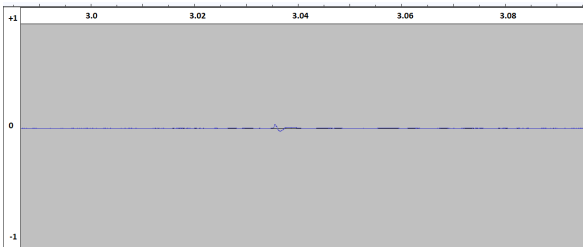
(a) *Sound wave of recording at corner P1*



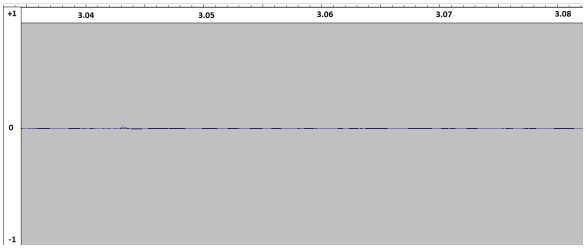
(b) *Sound wave of recording at corner P2*



(c) *Sound wave of recording at corner P3*

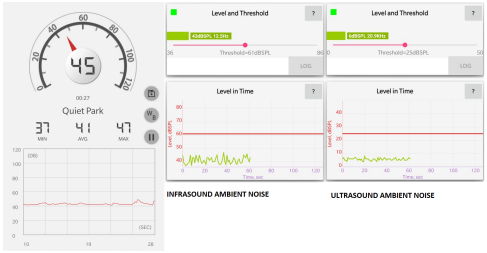


(d) *Sound wave of recording at corner P5*

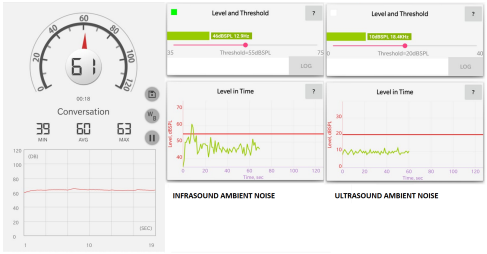


(e) *Sound wave of recording Outside*

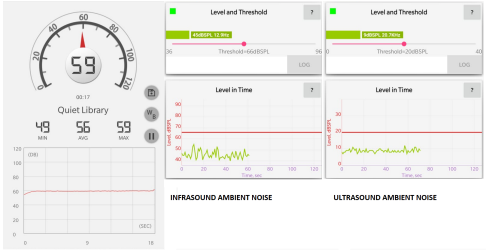
Figure 3.4: Sound waves of recordings taken at different positions and different locations



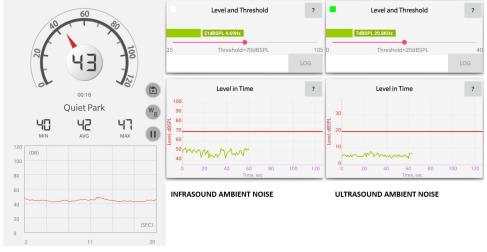
(a) *Ambient noise at corner P1*



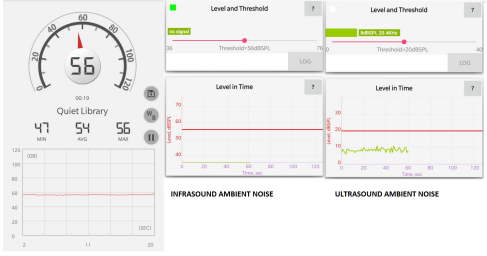
(b) *Ambient noise at corner P2*



(c) *Ambient noise at corner P3*



(d) *Ambient noise at corner P4*



(e) *Ambient noise at corner P5*

Figure 3.5: Ambient noise measurements taken at different positions in a room

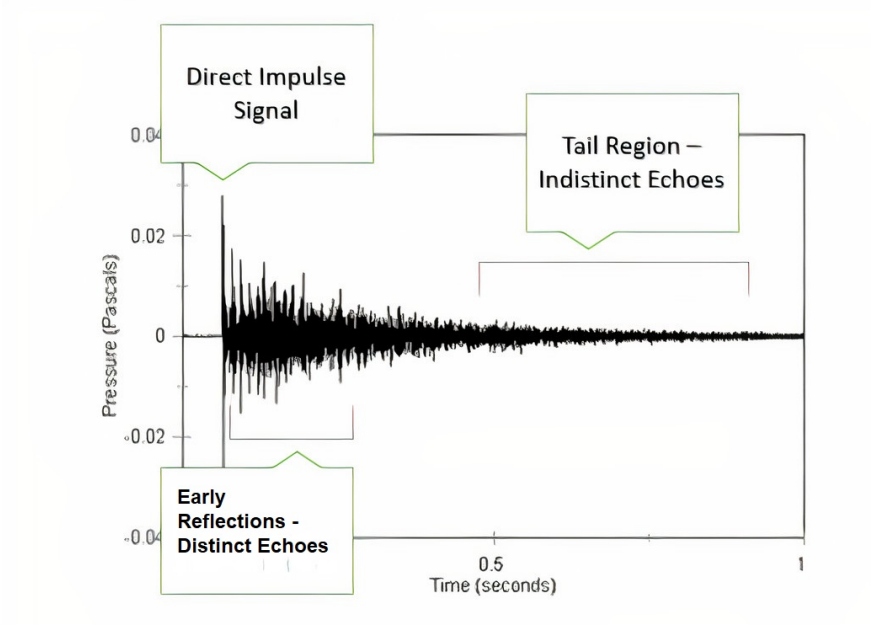


Figure 3.6: Impulse Response

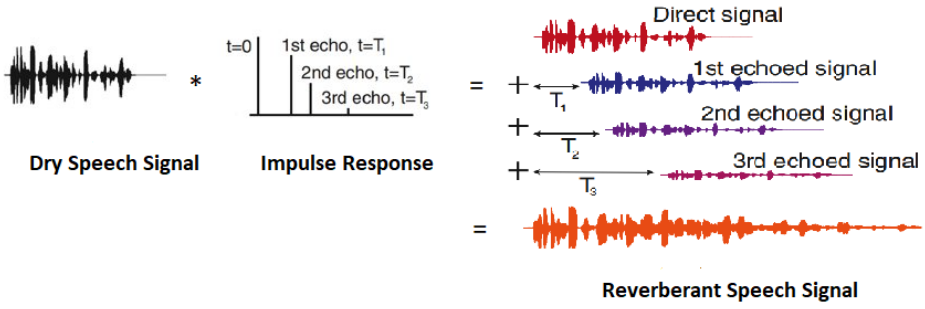


Figure 3.7: Reverberation in Speech Signal. Image borrowed from [26]

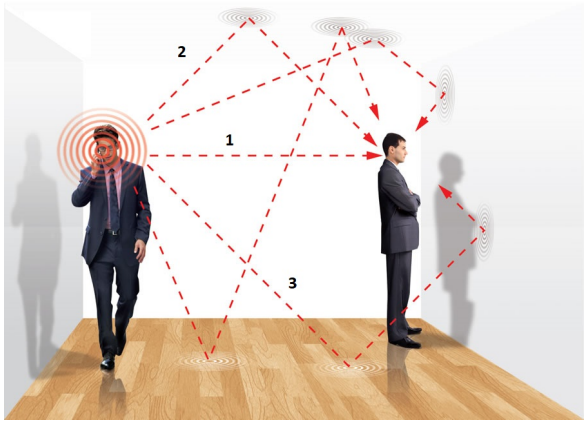


Figure 3.8: Example of reverberation in a room where path 1 refers to the direct sound wave and paths 2 and 3 refer to reflected sound waves

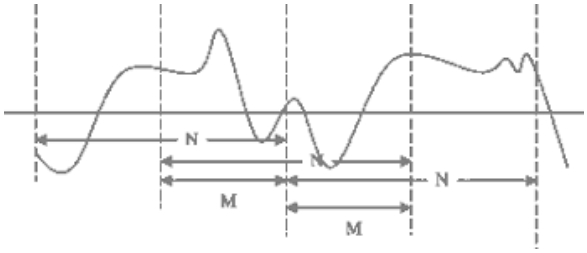
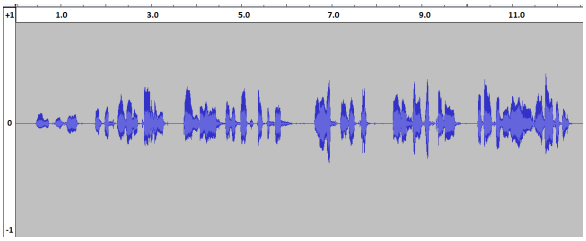
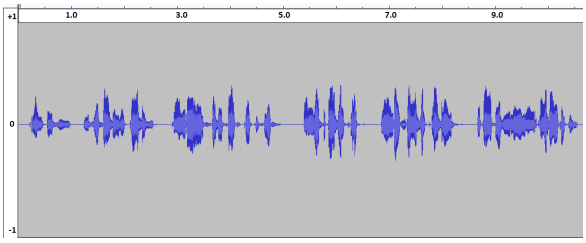


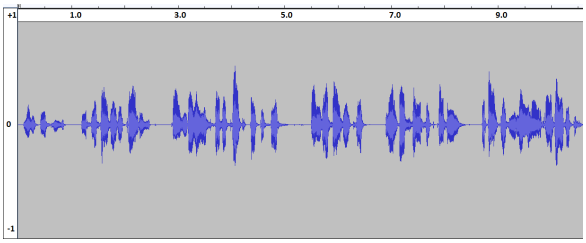
Figure 3.9: Segmenting a signal into frames. The image is taken from [45]



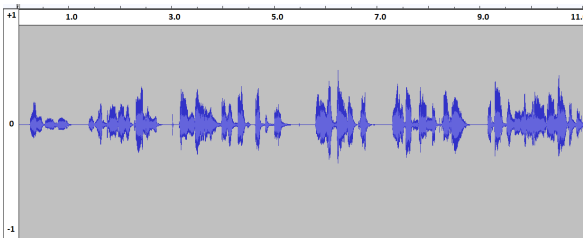
(a) *Sound wave of recording at corner P1*



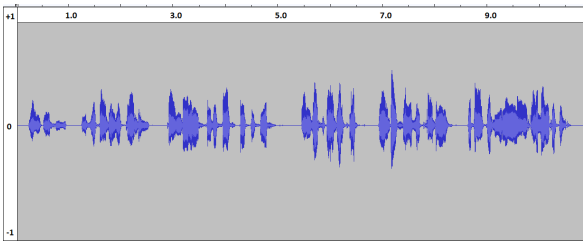
(b) *Sound wave of recording at corner P2*



(c) *Sound wave of recording at corner P3*



(d) *Sound wave of recording at corner P4*



(e) *Sound wave of recording at corner P5*

Figure 3.10: Sound waves of recordings taken at different positions in a room

4

MACHINE LEARNING CLASSIFIERS

In this chapter, we introduce some basic concepts used in classifier selection. Then we briefly describe each of the classifiers which we choose for our classification problem.

Classification is the problem of distinguishing to which category a given observation or set of observations belongs. The Machine Learning algorithms performing this task are known as classifiers. These classifiers can be categorized into various groups based on their learning techniques, problems they solve, etc. Based on the learning technique followed classifiers can belong to supervised learning algorithms, unsupervised learning algorithms, or semi-supervised learning algorithms. In supervised learning, the algorithm is provided with training data that includes the output variable (the class) corresponding to the input variable (a set of feature values). This allows the algorithm to then deduce the correlation between the input and output values so that it can infer the output value of a previously unknown input value (feature set). Some algorithms which come under this category are Linear Regression, Support Vector Machines, Logistic Regression, Linear Discriminant Analysis [4], k-Nearest Neighbor, Decision Trees, etc. In the case of unsupervised learning, the algorithm learns the behavior pattern of the input values without knowing the output value associated with them (unlabeled data). This task is much harder to realize when compared to supervised learning. Examples of such algorithms are k-means for clustering, Apriori algorithm for association rule learning. The third class of algorithms lies between the above-discussed classes, in the case of semi-supervised learning the algorithms have a combination of unlabeled and a couple of labeled input values. Most real-world problems tend to fall into this category due to the expensive and time-consuming nature of collecting and storing labeled data in comparison to unlabeled data.

Classifiers in Machine Learning can be grouped as generative or discriminative classifiers. While these classifiers have an identical end goal the means of achieving the same are different. In the case of generative classifiers, the model tries to represent the actual class distributions while with discriminative classifiers the model attempts to learn the difference between the classes by modeling the decision boundary or all the points where the classes are equally probable to occur. Some examples of generative classifiers are the Naive Bayes classifier, hidden Markov models, and Bayesian networks and few examples of discriminative models are Logistic Regression, traditional Neural Networks, and Nearest Neighbor.

4.1. AUDIO CLASSIFIER SELECTION

Deciding which Machine Learning algorithm works best for our dataset or problem is not a straightforward task. We often come across the *No free lunch* theorem in Machine Learning which states that no optimization algorithm performs better than the rest for all possible problems. Hence, we need to select what would work best for our scenario from among the available Machine Learning algorithms. Some of the factors that are of high importance during audio classifier selection are -

- **Available Training Data -** It goes without saying that the more the amount of data the better the model can perform as it has more data to learn from. However, the amount of data available is often constrained. This is also the case for our experiments as we shall see in Chapter 9. Since we have a good amount of features but lesser data points we favor algorithms such as Linear Regression which has a high bias but low variance.

- **Accuracy -** Machine Learning models can be flexible or restrictive. The restrictive model's output is highly interpretable as we can understand the relation between the predictor and the output. In the case of flexible models, we lose out on interpretability in the pursuit of better accuracy. For our problem, the accuracy or the ability of the model to accurately recognize the class is of utmost importance. Hence, for our scenario, a model with higher accuracy and more flexibility such as Support Vector Machines seem favorable.

- **Time -** The time or the speed with which the model *learns* is of high relevance for real-world applications. The time taken depends on the amount of training data and the classifier chosen. Typically higher accuracy means more time. For our scenario, since the data used is relatively small we concentrate on the choice of classifier and tuning its parameters such that we obtain fairly good accuracy and the implementation is fairly quick to execute.

4.2. CLASSIFIERS CONSIDERED

Keeping in mind the factors described in the previous section a subset of classifiers have been selected for our task. The selected classifiers are Linear Discriminant Analysis, Logistic Regression, Ridge classifier, Support Vector Machines, and Voting classifiers. The key idea and principle behind these classifiers are described here. We selected these models based on preliminary results we obtained using the *compare_models* function. This function trains every model present in the model library utilizing default hyperparameter values and then performs cross-validation. The resulting metric is averaged across all folds for every model and is displayed in a sorted tabular fashion. The default metric is accuracy which is also the evaluation metric we use for our classification task.

4.2.1. LINEAR DISCRIMINANT ANALYSIS

Linear Discriminant Analysis (LDA) is one of the commonly used classification techniques. In this method, the classifier tries to maximize the separability between classes so that a decision boundary between the classes can be drawn with ease. It assumes a multivariate Gaussian distribution of the data where the data is characterized by a mean and covariance and all the classes have an equal covariance value. LDA uses information from all the features in the model and creates a new axis. All the data points are then projected onto this new axis to provide maximum inter-class separability. Figure 4.1 shows how this works for a simple binary classification using two features.

While we have shown the process for a classification problem with just two features, the same can be extended to problems with more than two features. We retain the same process of creating an axis that maximizes the distance between two classes and minimizes their scatter. And so implicitly within LDA dimensionality reduction occurs one dimension at a time. Similarly, we can extend the process for problems with more than 2 classes. Now when we calculate the distance between the means of the classes we measure the distances between a main central point and the means of the classes. Also, we no longer have a single axis but multiple axes to separate the data. The process sounds very similar to that of principal component analysis for reducing dimensionality however they are different. In the case of PCA, the first new axis created will account for the

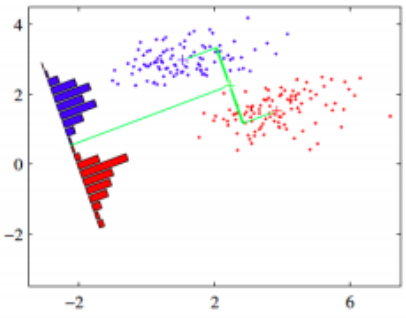


Figure 4.1: Projection of datapoints on new axis. The image is taken from [46]

most variation in data (feature with the most variation) while in the case of LDA the first new axis caters to the variation between the classes.

The new axis is created keeping in mind 2 criteria. These two equally important criteria are -

- The distance between the means of the classes must be maximized.
- The intra-class variation (also referred to as scatter) must be minimized.

These two criteria must be maintained simultaneously, otherwise it will lead to an overlap between the data points of different classes. The decision boundary is given by all the points where the probability of a point on the boundary belonging to either of the classes is equal. Hence using the Bayes rule we get the following equation:

$$P(y=k|x) = \frac{P(x|y=k) + P(y=k)}{P(x)} \tag{4.1}$$

In this equation we can rewrite P(x|y) as we have assumed a multivariate Gaussian distribution with equal covariances across classes:

$$P(x|y=k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\sum|^{\frac{1}{2}}} exp(-\frac{1}{2}(x-\mu_k)^t \sum^{-1}(x-\mu_k)) \tag{4.2}$$

After applying log to the equation 4.1 and replacing the value of P(x|y) with 4.2 we get the resulting equation:

$$logP(x|y=k) = -\frac{1}{2}(x-\mu_k)^t \sum^{-1}(x-\mu_k) + logP(y=k) + C \tag{4.3}$$

here C refers to the constant term from P(x). The colored term in the equation 4.3 refers to the Mahalanobis distance which accounts for the distance of a point from the mean of the classes and for the variance. From this equation, it is very obvious that the boundary surface is linear in nature.

4.2.2. LOGISTIC REGRESSION

Linear Regression: In this model, a linear relationship is assumed to be present between the input values (x) and the output value (y). In the case of a simple model, we fit a line to the data given to us as the value of y is a linear combination of the input variable x. We could then use this line to predict values of y given some other input values. If we increase the number of input variables we end up in a higher dimension and the "line" becomes a plane or a hyperplane.

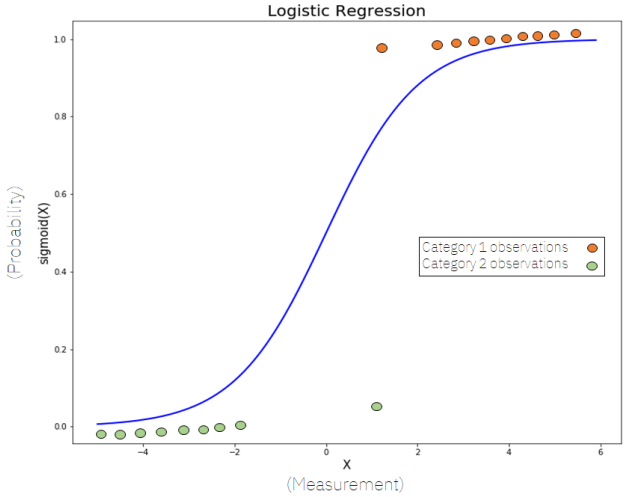


Figure 4.2: Fitting a sigmoid function on the given data. This image is taken from [47]

While a classifier labels a data point we may also be interested in ascertaining how sure the classifier is of its prediction. This is typically important in domains such as medicine, security, etc. Hence, one such method of achieving this is to treat the classification problem as a regression problem. Normally in classification, we tend to have labels, to convert it to a regression problem we use indicator variables. Linear Regression has certain problems regarding output interpretability as the output values are not within a limited range and are continuous in nature. So instead of directly using Linear Regression, regression is rather used on a transformed function. Now instead of fitting a line, we fit an "S" shaped function going from 0 to 1. The transformation used here is the *logistic* function or *logit* function. If p(x) denotes $p(y = \frac{1}{x})$ then the logit transformation is given by

$$log(\frac{p(x)}{(1-p(x))})$$

or the log of the ratio of the probability of success to failure. Hence, in Logistic Regression, we try to fit a Linear Regression model to the logistic function. Here the logistic function can be modeled as some linear function

$$log\frac{p(x)}{(1-p(x))} = \beta_0 + x * \beta_1 \tag{4.4}$$

On solving for p(x) we end up with p(x) looking like a sigmoid function as shown in Figure 4.2

$$p(x) = \frac{e^{\beta_0 + x*\beta_1}}{(1 + e^{(\beta_0 + x*\beta_1)})} = \frac{1}{(1 + e^{-(\beta_0 + x*\beta_1)})} \tag{4.5}$$

Varying β_1 changes the slope of this function while varying β_0, we obtain where the function is going to rise. The values run from 0 to 1, so it's a very valid means of fitting probabilities. Hence this solves the issue we had with using Linear Regression directly. This results in a very powerful classifier that can be used in varying domains and various settings. The decision boundary of this classifier is still given by a line which is

$$\beta_0 + x * \beta_1$$

4

so it is also a Linear classifier. Linear Regression also tends to make mistakes closer to the boundary in comparison to Logistic Regression due to the fact that we can have a steep climb from 0 to 1. So far the model was discussed with respect to two-class classification problems due to its ease. However, the same can be extended to k classes. Then each class gets a set of parameters β_0 and β_1. So, now the probability that the data point belongs to a certain class is given by

$$p(Y = c|x) = \frac{e^{\beta_0^{(c)} + x*\beta_1^{(c)}}}{\sum_l e^{\beta_0^{(l)} + x*\beta_1^{(l)}}} \tag{4.6}$$

The division by the sum acts as a normalizing factor. The main difference in Logistic and Linear Regression - with Linear Regression the line is fit on the data using least squares. We find the line such that the sum of the squares of the distance of the data points to this line is minimized. In the case of Logistic Regression, we use something known as maximum likelihood. Once the curve is fit we calculate the likelihood of the data for this particular curve. This curve is then shifted along and the likelihood (it is the product of the likelihood of all the data points) is calculated. Finally, the curve resulting in the maximum likelihood is chosen.

4.2.3. RIDGE CLASSIFIER

Bias: The ineffectiveness of a model in obtaining the true relationship of the data leading to underfitting.The bias is because the model oversimplifies the problem and doesn't give much importance to the training data.

Variance: It refers to the difference in fit between the test set and the training set.Or in other terms how differently the target function varies on changing the training dataset.This occurs when the model tries to overfit the training data and now performs poorly when confronted with new test data.

Bias-Variance tradeoff: If we have an overly simple model we end up with high bias and low variance but an overly accurate fitting model leads to a model with low bias but high variance. In Machine Learning we wish to achieve a model with optimal values of bias and variance which lies between these two extremes of overfitting and underfitting. We wish to achieve a model with both low variance and low bias.

The Ridge classifier like Logistic Regression attempts to perform classification using a regression method. Here the model converts the label of the data to [-1,1] and then performs the regression task in place of classification. Then based on the result of the regression model if the value is greater than 0 then the data point belongs to the target class mapped to +1, else it belongs to the target class which is mapped to -1. If there are more than two classes in the classification problem multi-output regression is used. Here multiple independent regression models are used with one model per class. Based on the predictions obtained from these models the class is predicted based on the highest prediction value obtained.

This classifier is based on the Ridge regressor. The highlight of Ridge Regression is that the model prevents overfitting the training data and creating a high variance model. To do this some amount of bias is introduced to create a slightly worse fit but this would result in better predictions of the model. In Linear Regression, the parameters of the line fit to the data are obtained by minimizing the least-squares value (sum of the squares of the minimum distance from the data points to the line).In the case of Ridge Regression, it minimizes not only the least-squares values but also λ*the square of the slope. This addition adds a penalty to the original method. The value of this lambda can range from 0 to positive infinity. A zero value of lambda results in 0 penalties and is the same as the least-squares or the Linear Regression. Increasing λ decreases the slope making it closer and closer to 0. So the optimal value of lambda can be chosen using cross-validation.

Considering a simple problem with one input and prediction one output as shown in 4.3 the change in slope shows how the relationship varies. In this figure, the red line refers to the line fit using least squares while the blue line is the line fit using Ridge Regression. A steep slope means the output is highly sensitive to changes in input while a lower slope means the opposite. We also notice how the line fit using Linear Regression has a higher slope compared to that fit using Ridge Regression due to the penalty term.

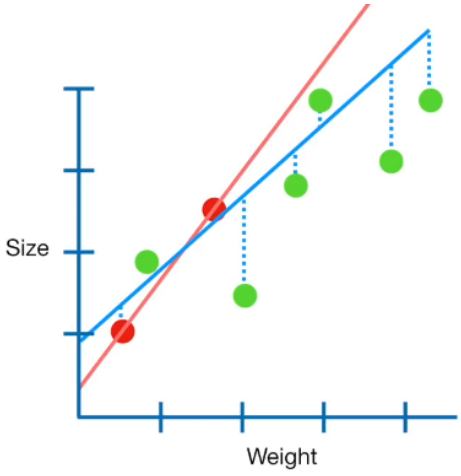


Figure 4.3: Fitting the datpoints using Linear and Ridge Regression. This image is taken from [48]

4.2.4. SUPPORT VECTOR MACHINE

SVM is a linear model which can be applied to solve linear and non-linear problems (datasets that are non-linearly separable). Since it is a linear model, SVM finds the line (plane or hyperplanes in higher dimensions) separating the classes. However there need not be a unique line or hyperplane performing this task, so SVM identifies the best line (hyperplane) for the given problem.

Figure 4.4 shows three of the multiple possible lines that separate the data points of the different classes. However, the line that performs best is the line in the middle of the widest street separating the classes. The data points that are closest to the hyperplane (the circles and triangles that are filled in the figure) are known as support vectors. The shortest distance between these support vectors and the threshold (optimal hyperplane) is called the margin. The underlying principle of Support Vector classifiers is to maximize the separability between the classes. This results in a maximal margin classifier. The margins shown in the figure are hard margins.

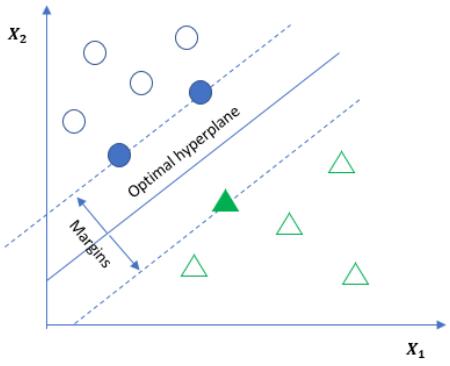


Figure 4.4: Optimal line separating the classes using Support Vector classifiers. This image is taken from [49]

An issue with maximal margin classifiers is their sensitivity to outliers. So to resist this behavior we choose a threshold that allows for misclassification. This is an example of the bias-variance tradeoff. When we choose such thresholds the resulting margins are called soft margins. There can be data points that lie within these soft margins unlike the case with support vectors lying on the hard margins. These data points which lie within and on the soft margins are called support vectors and hence the term support vector classifiers arise. We again end up with the issue that there is no unique soft margin, so to choose which soft margin performs the best we can use cross-validation. The model can be extended to data in the higher dimensions too and instead of a line a plane or a hyperplane separates the classes. So for an n-dimensional problem, an (n-1) dimensional hyperplane is created.

So far we have made the assumption that the classes are separable. However, Support Vector Machines also take care of problems where this is not the case. SVM implicitly projects the datapoints onto a higher dimensional space where the classes are separable and then creates the soft margins like the support vector classifiers and performs classification. The data in a higher dimension is calculated with the help of kernel functions

but no actual data transformation is performed. *The kernel trick* is the calculation of high dimensional relationships without performing an actual transformation. There are multiple kernel functions that can be used for this task such as polynomial kernel function, radial kernel function, etc.

So far we have discussed how classification is performed when we have binary classification problems. SVM can also be used when we have multiple classes. The principle used here is to break down the multi-class classification problem into multiple binary classification problems.This is achieved in two ways:

- ***One-to-One Approach:*** This approach takes into account only 2 classes at a time while ignoring the remaining classes. So the SVM generates a hyperplane that separates the datapoints of the classes considered.This results in a total of $\frac{\mathbf{n}(\mathbf{n}-1)}{2}$ SVMs for a **n** class problem.

- ***One-to-Rest Approach:*** In this approach an SVM is used per class so for a **n** class problem **n** number of SVMs are used.The SVM will create a hyperplane that separates a single class from the data points of all the remaining classes. In this scenario, the two classes considered for binary classification are one group that contains the datapoints of the class and the other group containing the datapoints of all the remaining classes.

4.2.5. VOTING CLASSIFIER

Voting classifiers are Machine Learning models that perform classification based on the outputs obtained from all the contributing classifiers present in their ensemble. The way in which the output of the different Machine Learning algorithms is combined is by voting as the name suggests. However, this voting can be done in two ways-

- ***Hard Voting:*** Each of the contributing classifiers predicts the class label for the test data. Then the voting classifier checks to see the majority outputs of these contributing classifiers corresponding to the labels. This label is then taken to be the output of the voting classifier. In case of no clear majority, the class is chosen based on the ascending sort order.

- ***Soft Voting:*** Instead of performing voting in a traditional sense here, a weighted average is used. Here each of the classifiers in the ensemble is assigned a weight. When performing classification the classifiers assign probabilities with which the datapoint belongs to all the possible classes. The soft voting classifier then performs a weighted average of all these probabilities. Then the Voting classifier chooses its output based on the class having the highest average probability.

5

Audio Features and Audio Analysis

In this chapter, we address the issue of how to select features for audio classification. We also describe the audio features we have utilized in our classification model.

5.1. AUDIO FEATURES CLASSIFICATION

An audio signal is a three-dimensional signal with each dimension representing amplitude, time, and frequency respectively. Features of this signal characterize a raw sound wave and can be categorized as physical features and perceptual features. Perceptual features refer to features that are characterized by the way a human perceives sound and these features were discussed in 3.1. Physical features are those that can be computed from audio mathematically. Such kinds of features can further be grouped based on the representation domain such as Temporal (Time domain) features, Cepstral features, frequency domain features, and frequency-perceptual domain features etc[16].In the time domain, the signal is represented as amplitude vs. time, and in the frequency domain, the signal is shown as amplitude vs. frequency.

Features can be extracted at two levels, short term level or frame level and a long term level. Traditional speech processing mainly uses frame-level features. An audio wave is divided into many frames and features can be extracted from each of these frames to obtain frame-level features. These features help to represent any short-term behavior of the audio signal. For our research problem, we opt to use frame-level features based on previous literature[13]. How the features are selected and a brief description of the selected features are discussed in the following sections.

5.2. FEATURE SELECTION

Choosing a lot of features means more data, which means more information and better machine learning models right? No, this is not always the case.Instead, it is highly imperative to choose the right features from amongst the features that can be irrelevant and insignificant.If the significant features are not selected we end up with an inefficient model which utilizes unnecessary resources.Features are selected based on certain considerations which are discussed below:

- ***Size:*** This factor is a direct result of *the curse of dimensionality*[6]. With the increase in the size of data, there is an increase in the error. Hence, it is important to ensure there are just enough amount of features to completely characterize the sound waves without introducing errors. Using an optimal feature size leads to simple models that perform well.
- ***Complexity:*** This refers to two things, the time and resources taken to calculate feature values and the resources taken by a model to perform its task based on the features selected. Ideally, we wish the time and resources used to be kept at the bare minimum. Some features require more time to be computed than others. Also, the complexity of the model itself is a byproduct of the size of the feature set - a bigger feature set requires more resources computationally.
- ***Class behavior:***
  - ***Intra-class behavior:*** Low variation in feature values within a class is required. The similarity in feature values between data points will then help to group them together as part of the same class. This helps achieve low variability which is highly desired in classification problems.

  - ***Inter-class behavior:*** High inter-class separability or high variability between classes is another behavior that is highly necessary for classification problems. Hence it is optimal if the features selected have highly different values across classes. This will further help distinguish data points showing different behavior as belonging to two separate classes with ease.

- ***Sensitivity:*** Data is always assumed to have some amount of noise, so slight variations in the input should not affect our feature values. It is evident that a highly sensitive set of features results in a model that is not powerful enough to perform well in case of noise.

- ***Independent:*** The features should be independent and not correlated to any other feature. It is a good practice to remove any dependant features as they are redundant. If we have a couple of dependant features in our dataset then the classifier is going to give a lot of prominence to that feature introducing bias. For instance, having a feature weight in kg and another feature weight in pounds is redundant and the model considers both of them separately, which increases the importance that *weight* has in this model.

5.3. AUDIO FEATURES

Since the task of filtering out the best features for a problem is not so easy we have many machine learning techniques used to do the same. Some of these techniques are forward selection, backward elimination, recursive feature elimination, etc. However, since there are previous works in this field we leverage their findings and base our features on them.Based on the literature works the following features were selected - Zero Crossing Rate[21], Spectral Roll-off[5], Spectral Flatness, Spectral Centroid, and Mel-frequency cepstral coefficients (MFCC)[2][19].We briefly describe these features in the following subsections.

5.3.1. ZERO CROSSING RATE

This is one of the simplest and easiest features to calculate computationally. It is a temporal feature that refers to the rate at which the signal moves from positive to negative or back or in other words, it's the rate at which the signal changes signs. This feature is especially useful to distinguish unvoiced speech or speech that doesn't make use of the vocal cords and sounds due to percussion. We can calculate this ZCR value by keeping a count on the number of times the signal crosses the zero axis. Mathematically ZCR[21] is given by the following equation -

$$ZCR = \frac{1}{2W_L} \sum_{n=1}^{W_L} |sgn[x_i(n)] - sgn[x_i(n-1)| \tag{5.1}$$

here the sign function is denoted as sgn(.),it takes a value of 1 when the signal value x_i(n) is greater than or equal to 0 and a value of -1 when the signal x_i(n) is less than 0. The value W_L refers to the length of the window under consideration. ZCR is also indicative of the amount of noise in a signal, a higher ZCR value typically means more noise. In the figure 5.1 we see that for the signal depicted we can count three zero crossings for the duration of the signal from 0 to 100-time units.

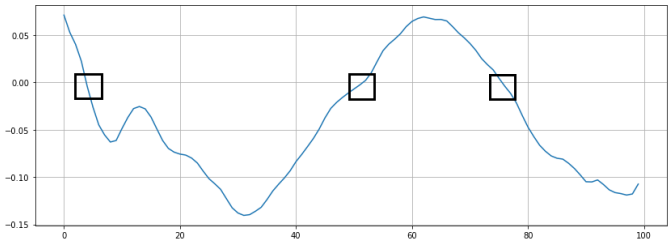


Figure 5.1: Zero Crossings in a signal

5.3.2. SPECTRAL ROLL-OFF

Spectral Rolloff refers to the frequency below which a certain percentage of the spectral energy is contained. The default percentage of energy considered is 85%. This frequency is calculated for every frame in the signal and the energy of the spectrogram in this frame must be contained by this frequency. It is especially useful to differentiate harmonic sounds from noisy sounds as noisy sounds usually lie above the roll-off frequency. The figure 5.2 depicts the signal waveform in blue and the roll-off frequencies calculated for every signal frame in red color.

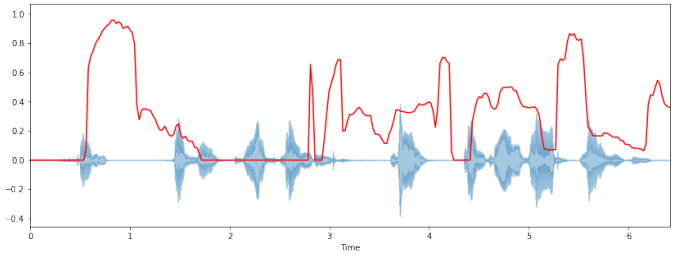


Figure 5.2: Spectral Rolloff

5.3.3. SPECTRAL FLATNESS

Spectral flatness a.k.a Wiener entropy is usually measured in decibels and is used as a means of quantifying how tonal vs how noisy a sound is. Mathematically this value is calculated as the ratio between the geometric and arithmetic means of a power spectrum. The power spectrum represents how the power is distributed among the different constituent frequencies of a signal. Formally we can derive the spectral flatness value as -

$$Spectral\_Flatness = \frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\frac{\sum_{n=0}^{N-1} x(n)}{N}} = \frac{exp(\frac{1}{N}\sum_{n=0}^{N-1} \ln x(n))}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)} \tag{5.2}$$

Here x(n) represents the magnitude of the frame or bin. This ratio is then converted and reported on the dB scale.

5.3.4. SPECTRAL CENTROID

Geometrically the centroid helps to obtain the center point of an object and for simple uniform objects, this point also refers to its center of mass. The spectral centroid performs a similar function with respect to a spectrogram. Mathematically this value is the weighted mean of the constituent frequencies of a signal, which is determined using the Fourier transform as shown below -

$$Spectral\_Centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \tag{5.3}$$

The weights in this weighted mean are obtained from the magnitudes. x(n) depicts the weighted frequency value, n represents the bin or frame number, and f(n) refers to the center frequency of that frame. The spectral centroid of the signal is shown below in 5.3, the waveform is depicted in blue and the spectral_centroids in red.

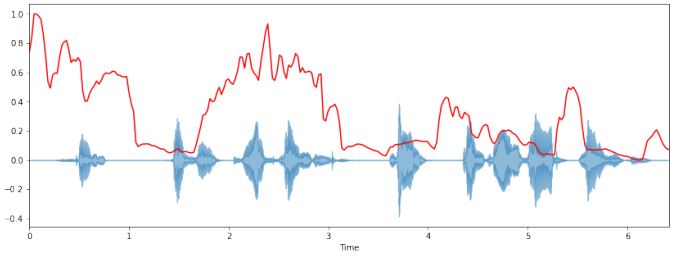


Figure 5.3: Spectral Centroid

5.3.5. MFCC

This feature is synonymous with speech and voice recognition. This set of coefficients help depict an all-inclusive shape of the spectral envelope giving details regarding timber. MFCCs are considered very powerful when considering audio with machine learning as it takes into account the non-linear behavior of the human auditory system with respect to different frequencies. By default, for the librosa package (version 0.8.1) in python, the number of MFCC's calculated is 20.

An audio wave can be depicted either in a time domain or a frequency domain and we can also perform transformations to convert a signal between these domains. The Fourier transform is a means of converting a time-domain signal to that of the frequency domain. On performing a Fourier transform on the signal in the time domain we obtain a Fourier spectrum which is used for MFCC extraction. Some amount of preprocessing is performed initially involving pre-emphasizing and windowing. Pre-emphasizing is to perform noise reduction and windowing (hamming window) is used to avoid leakages due to discontinuities present at the edges with the help of smooth functions. The spectrum is then converted to mel-scale using a filter. The mel-filter bank helps to perceive the audio like our human auditory system. Humans are more receptive to small changes at lower frequencies and tend to have a non-linear perception of sound. Then redundant information is removed by de-correlation using the discrete cosine transformation and

the log steps. The process to obtain the MFCCs is to then take a log of the magnitudes of the filtered Fourier spectrum, doing this helps to retain amplitude information while disregarding phase information which is considered not very relevant. How loud a human hears a sound is approximately on the logarithmic scale, hence we perform a log on the magnitude and then a DCT on them. The process of MFCC extraction is shown in 5.4. Performing these steps results in a spectrum that belongs to neither the frequency

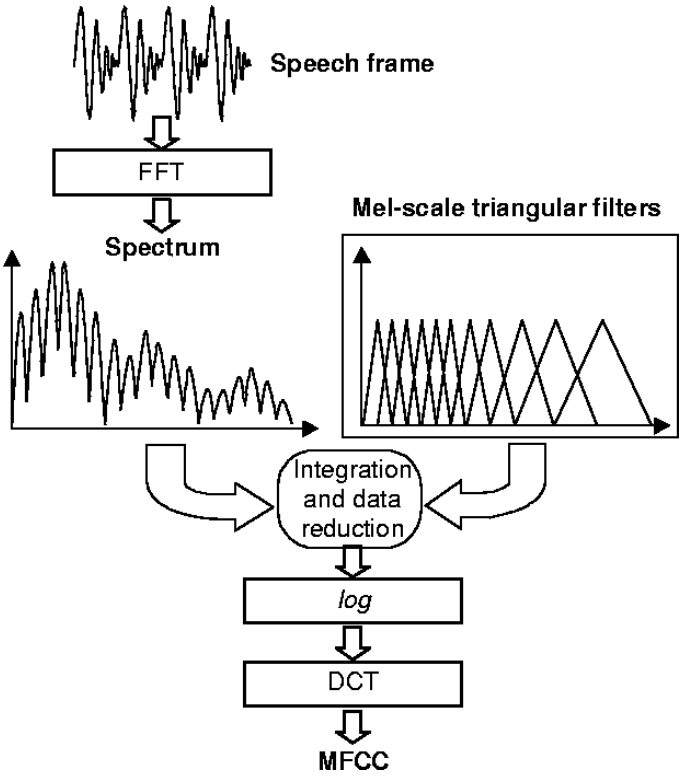


Figure 5.4: MFCC Feature extraction. This image is borrowed from [7]

nor the time domain. The spectrum of the log of the Fourier spectrum is termed as a *cepstrum*.

6

System Adversary Model

In this chapter, we describe the system and the adversarial model of our attack. We provide insight into the assumptions we make regarding the attacker and the victim. We further discuss the different types of realistic attack scenarios that we identified based on varying levels of information available to the attacker.

6.1. System Model

We assume that the victim has a smartphone device with WhatsApp installed and an internet connection. We further assume that the software on the victim device and the device itself is not compromised in any manner. Also, the victim sends the attacker audio messages via Whatsapp. While recording the audio messages, we assume that the phone is held at a distance of approximately 15 cm (which is well within the **critical distance**[1]) from the face of the speaker at an upright position as shown in Figure 6.1. This is one of the most common positions in which a phone is held either during video calls or while sending audio messages.

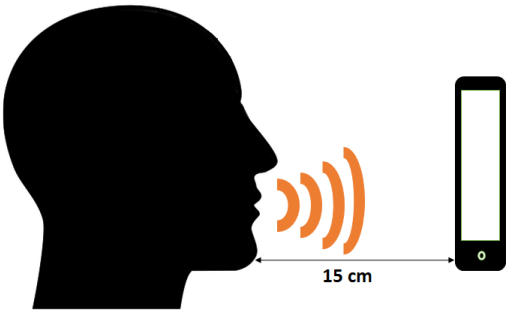


Figure 6.1: Recording position

6.2. Adversarial Model

We assume that the attacker has access to the WhatsApp audio message of the victim. The attacker is a user who seeks to learn the location information of the victim. Here the term **attacker** refers to even investigators in case of a forensic investigation and inquiries and not an attacker in the traditional sense. Depending on the attack scenario, the attacker may be assumed to also have the target's recordings from the same or different positions at specific locations. Also, the victim is assumed to be in one of these selected locations at the time of recording the audio message. For our experiments we consider three different scenarios for the attacker:

- **Complete Profiling**: This scenario occurs when the attacker asks the victim to send voice messages from specific locations. For example, an investigator (i.e., the attacker) might ask a suspect (i.e., the victim) to stand in a specific part of a room to verify that at the time a voice message was sent, the suspect was there or elsewhere. In this scenario, the attacker has recordings of the victim in all the selected locations. Moreover, the attacker knows also the specific position of the victim in the selected locations (e.g., in a corner of a room). In this scenario, the attacker has the highest knowledge to execute his attack as both the victim and the location are "known" in the training set.

[1] Critical distance is the distance between the microphone and the sound source at which the level of room reverberation is same as the level of the direct sound.

- **Location Profiling**: In this scenario, the attacker cannot access any of the victim's voice messages, other than the one for which he wants to infer the location. The attacker knows that the victim has sent the voice message from one of the selected locations (e.g., the attacker knows that the victim is in a specific building). Therefore, the attacker can have WhatsApp audio recordings of different speakers but the victim. The speakers are assumed to have recorded their messages at the same location positions from which the victim is sending the voice message. Hence, the victim is "unknown" while the location position is "known" to the attacker.

- **User Profiling**: This scenario occurs when the attacker owns the victim's voice messages, he knows the location they were sent from but does not know the specific position in the location (e.g., a corner of a room) from which they were recorded. The attacker wants to infer the location of a new voice message sent by the victim. Different from the *Complete Profiling* scenario, the attacker cannot ask the victim to send more voice messages from specific positions of the selected locations (e.g., the victim is no longer reachable). In this situation, the victim is "known" while the position is "unknown" to the attacker.

7

For Your Voice Only ATTACK

In this chapter, we provide an overview of the different phases that are a part of our For Your Voice Only attack. We then describe each of the four phases in detail.

Our attack consists of four different phases: Data Acquisition, Data Processing, Model Training, and Location Inference. In Figure 7.1 we provide an overview of how the attacker conducts the attack. Each of the four phases is discussed in detail in the following sections

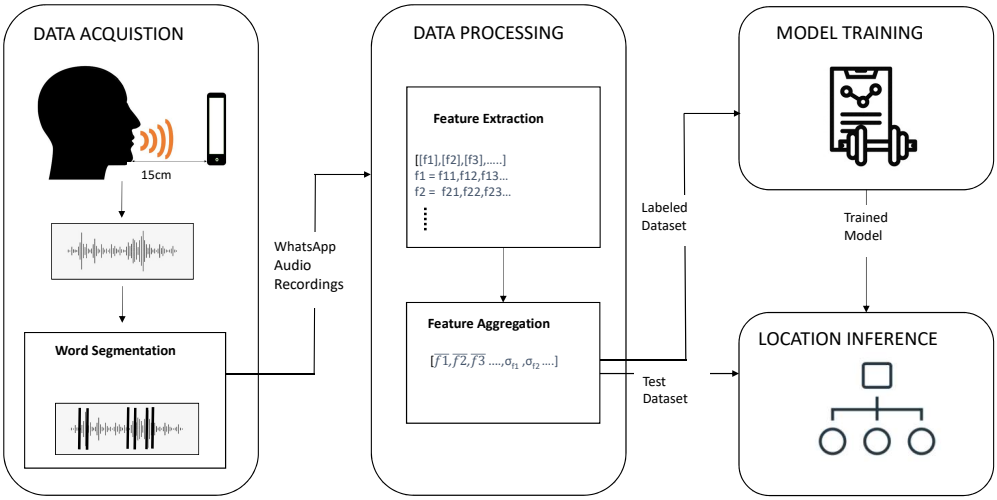


Figure 7.1: *ForYourVoiceOnly* attack phases

7

7.1. DATA ACQUISITION

This phase consists of two steps: Recording and Word Segmentation. At the end of the data acquisition phase, the attacker will be in possession of two datasets composed of segmented voice messages. The obtained data is in the OGG file format which is a lossy compressed file format not commonly supported by python libraries, so we convert our files to WAV format.

- *Recording*:The attacker selects his target victim and performs reconnaissance to select the locations of interest. In this step, the attacker performs two types of data acquisition. The first involves acquiring WhatsApp voice messages recorded by different people (including the victim if allowed by the attack scenario) at some locations or specific positions of interest to build a labeled dataset. The second, for acquiring unlabeled (i.e., both the location or the position are unknown) WhatsApp audio messages of the victim (i.e., test dataset). These two steps do not necessarily have to be consecutive. The attacker can create the labeled dataset even after obtaining the test dataset. The attacker can then choose the locations of interest, based on the type of information available (e.g., the victim might say she is in one location, but the attacker suspects she is in another known specific location).

- *Word Segmentation*: The attacker segments the recorded voice messages to extract audio fragments related to specific words frequently used in speech [15, 17] (e.g., "*and*", "*of*" and "*the*"). This procedure can be done either manually or by using speech-to-text algorithms. [1]

7.2. DATA PROCESSING

The data processing phase is carried out on both the labeled and the test datasets. This phase consists of two stages: *Feature Extraction* and *Feature Aggregation*.

- *Feature Extraction*: The attacker extracts frame-level features that are descriptive of vocal and environmental characteristics: spectral centroid, spectral roll-off, spectral flatness, zero-crossing rate, and Mel-frequency cepstral coefficients [19]. At the end of this step, the attacker has a set of time-frequency features whose dimensionality depends on the duration of the segmented voice message.
- *Feature Aggregation*: Since segmented voice messages may have a variable duration, the attacker needs to process the feature extracted in the previous step to create a feature vector of standardized length. The attacker aggregates the extracted features by calculating the average and the standard deviation as suggested in [10, 34]. This procedure allows maintaining information about the magnitude and variability of the data, reducing the total number of features per voice message. At the end of this step, each segmented voice message has a set of 48 associated features.

7.3. MODEL TRAINING

In this phase, the attacker uses only the labeled dataset. The attacker trains classification models. The attacker may also decide to train the models using a sub-sample of the dataset based on the information he owns. For example, in the acquisition phase, the labeled dataset may contain records from many locations, but the attacker has obtained new information about the victim and may discard some of them.

7.4. LOCATION INFERENCE

In this phase, the attacker applies the model trained in the *Model Training* phase and predicts the location or the specific location where the message was recorded by the victim. The input and output of this phase is reliant on the attack scenario considered:

- In *Complete Profiling* and *Location Profiling* the training set data available to the attacker contains information on the victim's location. The attacker has already trained the model with this information. Now the attacker tries to classify the target location from one of the identified locations. He further identifies the location position of the attacker while sending the audio message.
- In *User profiling* the attacker has no knowledge of the location position. Hence, here the attacker tries to simply identify the location of the target.

[1] https://www.mathworks.com/help/audio/ug/audio-labeler-walkthrough.html

8

PRELIMINARY EXPERIMENTS

In this chapter, we describe all the experiments carried out to help formulate the setup of our final experiment. These experiments are mainly divided into 4 sections - the first two based on device position and device model and the latter two based on the message content and the use of speakers. The detailed results can be found in the Appendix B.

8.1. DIFFERENT POSITIONS

In our final experiment, we carried out data collection using WhatsApp audio messages recorded by the participants while holding their phone at a distance of about 15 cm away from their face as shown in Figure 6.1. This position was based on one of the most common ways in which the phone is held especially during video calls and audio messaging. However, another highly used position is holding the phone against the ear when we receive audio calls as shown in 8.1. We collected 460 audio samples from a reduced pool size of 3 locations (i.e., two studio rooms and one outdoor location-balcony) for both the positions as shown in 6.1 (position 2) and 8.1 (position 1). We

Figure 8.1: Phone position against the ear (position 1)

noted that $\mathscr{F}orYourVoiceOnly$ reached an average accuracy of nearly 100% in predicting between the outdoor location and any one of the indoor locations with both the positions. Further, when trying to classify between all the three locations our attack resulted in an accuracy of 98% for position 1 and 97% for position 2. These results demonstrate that $\mathscr{F}orYourVoiceOnly$ can be applied even when the audio samples are collected from audio phone call recordings. There is only a decrease in the accuracy of $\mathscr{F}orYourVoiceOnly$ across these positions when performing corner classifications as seen in 8.2 which was expected. This is because in position 1 when the phone is held against the ear we expect that not all signals are picked up by the phone microphone unlike in position 2, as some of the waves may be obstructed by the speaker. The reported results are based on the classifier which performs the best - which is usually the soft voting classifier.

8.2. DIFFERENT PHONES

In our final experimental setting, we carried out data collection by having the participants record the audio messages on their own phones. This can result in a bias for certain members who have phones with better microphone characteristics. To test how the usage of different phones affect the attack scenario we collected 575 audio samples from a reduced pool size of 3 locations (i.e., two bedrooms and one outdoor location-terrace) where set 1 consisted of 345 samples which were recorded by participants on their own devices (OnePlus 3, OnePlus Nord and OnePlus 6T) and the second set contains 345 samples that were recorded on a common device, a Oneplus 3 handset (115

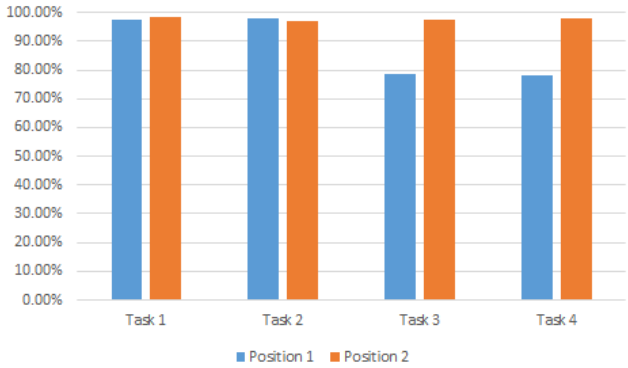


Figure 8.2: Accuracy of $\mathcal{F}orYourVoiceOnly$ with different phone positions. Here task 1 is room classification between all 3 locations, task 2 is room classification between 2 indoor rooms, task 3 is position identification between all 3 locations, and task 4 corresponds to position identification between both indoor rooms.

samples from the first set were reused here as the device was the same - OnePlus3 handset). Since we used the lower end model as our device for recording data in the second set, we expect the data obtained here to perform either worse or the same as the data obtained from different phone models (data in set 1).

We noted that $\mathcal{F}orYourVoiceOnly$ reached an average accuracy of nearly 100% in predicting between the outdoor location and any one of the bedrooms with both sets of data. When trying to classify between all the three locations our attack resulted in an accuracy of 99.6% with data in set 1 and 90.9% for data in set 2. Further, when performing position classification across these locations $\mathcal{F}orYourVoiceOnly$ achieves an accuracy of 71.2% with different phone data and accuracy of 59.8% with same phone audio recording samples. These results demonstrate that $\mathcal{F}orYourVoiceOnly$ is affected by the phone model used during recording. There is a decrease in the accuracy of $\mathcal{F}orYourVoiceOnly$ which was expected as mentioned previously. The reported results are based on the classifier which performs the best - which is usually the soft voting classifier. Additionally, the use of different phones is more realistic considering our attack scenario where we don't assume to have any prior knowledge about the victim's device. Also given the pandemic, this method of collecting data on their own handsets was preferred for our final experiment.

Further, we tried to assess if using the same phone for training data as used for the test data (victim) would help us in classifying test data better. For this purpose, we train our model using two speakers from our reduced pool size of 3 speakers (i.e., Speaker C, Speaker D, and Speaker E) and test our model on the data of the speaker that is left out (similar to leave one out validation so that all three speakers end up as test data in different iterations). A summary of the obtained accuracy for $\mathcal{F}orYourVoiceOnly$ is shown in table 8.1 with an explanation of the tasks in the caption. We expected the data obtained using a single device to help aid in the classification of new test data from unknown speakers. We notice that the indoor-outdoor classification accuracy is mainly affected by the different sets of data - the same phone data performs better or equally

well as audio captured from different handsets. While for the other tasks there seems to be no clear indication as to whether the use of the same phone for testing and training is really aiding in classification. The deviation in the results from what we expected can be due to the difference in microphone of the device as it is a lower-end model and its inability to pick up certain signals which aided in classification when compared to the other devices (OnePlus Nord and OnePlus 6T) used or simply the use of not enough data to make a conclusive inference.

Table 8.1: Accuracy of $\mathcal{F}orYourVoiceOnly$ based on device used. Here task 1 corresponds to indoor-outdoor classification, task 2 is room classification between all three locations, and task 3 is room classification between the two bedrooms.

Test Data	Speaker C		Speaker D		Speaker E	
	Different Phone	Same Phone	Different Phone	Same Phone	Different Phone	Same Phone
Task 1	89.2%	88.45%	90.8%	100%	99.25%	100%
Task 2	63.5%	56.5	64.3%	69.5%	86.1%	73%
Task 3	76%	50%	70%	70%	90%	75%

8.3. DIFFERENT AUDIO CONTENT

In our final experiment, we carried out data collection on WhatsApp audio messages recording single words pronounced by the participants. However, to study the impact that the content of the audio messages has on our $\mathcal{F}orYourVoiceOnly$ attack we collected different data sets comprising of audio recordings of varying content which are described below.

8.3.1. DIFFERENT MESSAGE CONTENT

In a real attack scenario, we have no control over the content of the audio messages that are sent. Hence, we perform a preliminary investigation to study the effect that different audio content has on the accuracy of $\mathcal{F}orYourVoiceOnly$. For this experiment, we collected 330 audio data samples of a single speaker and a single handset device over 4 locations (i.e., three bedrooms and one outdoor location-terrace). Of these 330 samples, 165 samples corresponded to the messages with the same statement as content while the remaining 165 samples consisted of different statements with no repeating audio content. We expect the audio content of the messages to contribute towards the classification ability of our models and the use of the same audio data may prove to be favorable but the use of different audio is not expected to completely eliminate the distinguishing ability of the models. This is due to the variation of characteristics of signals produced when pronouncing a certain syllable, so the presence of the same words in the training and test will prove to be useful. We base our hypothesis on the fact that $\mathcal{F}orYourVoiceOnly$ is dependant on the behavior of the signals during reverberation in addition to background noises which are still retained across different audio messages albeit with slight variations.

As shown in Figure 8.3 we note that $\mathcal{F}orYourVoiceOnly$ reached an average accuracy of nearly 99% in predicting between the outdoor location and any one of the indoor locations when using the same audio content in training and testing while the accuracy drops to 81.7% when the audio content is not the same. Further, when trying to classify

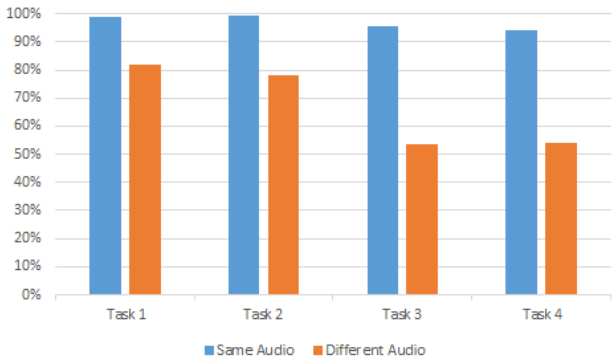


Figure 8.3: Accuracy of $\mathit{ForYourVoiceOnly}$ with different audio content. Here task 1 is room classification between all 3 locations, task 2 is room classification between 2 indoor rooms, task 3 is position identification between all 3 locations, and task 4 corresponds to position identification between both indoor rooms.

between all the three indoor locations our attack resulted in an accuracy of 99.2% for the same audio content data and 78% for different audio data. These results demonstrate that $\mathit{ForYourVoiceOnly}$ performs much better when the audio data content present in the training and testing data is the same. There is also a huge decrease in the accuracy of $\mathit{ForYourVoiceOnly}$ attack when performing corner/position classifications as seen in 8.3 which was expected. The reported results are based on the classifier which performs the best - which is usually the soft voting classifier.

8.3.2. SILENT AUDIO CONTENT

The presence of speech content in the audio recording samples no doubt aid in the classification of the recording locations of the audio samples especially in the classification of the positions. However, we revert back to our claim that the audio samples at different locations differ due to two reasons - the presence of ambient noise and reverberations. To show that the ambient noise differs at different locations and also different positions within a room, we conducted a preliminary study to see how $\mathit{ForYourVoiceOnly}$ attack works on silent data. For the purpose of this study, we used a total of 330 data samples of which 265 are new audio samples. These 265 samples comprise no speech content in the audio recording. To further see how the behavior of the model changes with more data we increase the amount of data available within a location from 50 to 250 (in steps of 50) to see whether this increase helps in better position classification within the room. If the performance of $\mathit{ForYourVoiceOnly}$ increases with data then more data means more useful information, this is the case when we are dealing with a high variance problem wherein the model is overfitting on the training data. However, if we have a simple model which is not effectively using the data leading to a high bias we won't benefit from this increase in data, in which case we have to reevaluate the models considered.

The $\mathit{ForYourVoiceOnly}$ attack performs really well with silent data when we try to classify indoor and outdoor locations as seen in the graph for task 1 in Figure 8.4. This result was expected due to the evident differences in ambient noise in both locations.

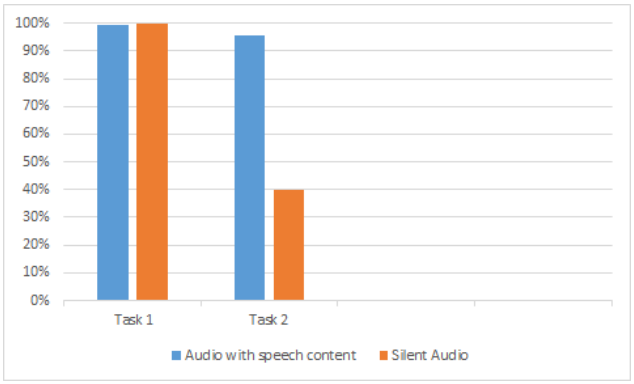


Figure 8.4: Accuracy of *ForYourVoiceOnly* with silent audio content. Here task 1 is indoor-outdoor classification and task 2 is position identification between both locations.

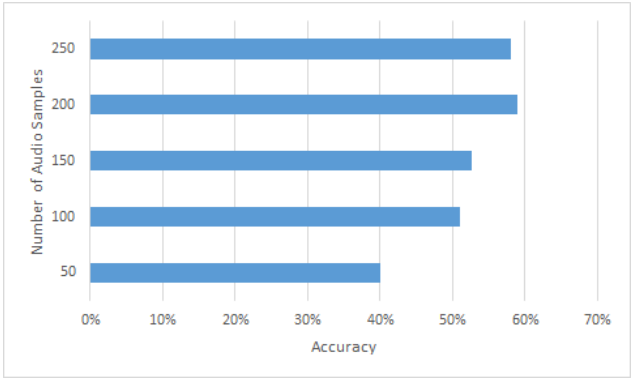


Figure 8.5: Accuracy of *ForYourVoiceOnly* for task 3 when we vary the number of silent audio samples. Here task 3 corresponds to position identification within an indoor room.

When the ambient noises are not so blatantly different the *ForYourVoiceOnly* attack faces difficulty in classification and the accuracy drops as seen with the bar for task 2 (position classification within the room) in the graph in Figure 8.4. We notice from Figure 8.5 that an increase in data points from 50 to 250 (in steps of 50) helps in increasing the accuracy of this task which means that we are dealing with a high variance problem and the additional information is proving to be useful.

8.3.3. SYLLABLES

In a real-world scenario, voice messages can have any content with no restrictions. Based on the results in 8.3.1 we see that our *ForYourVoiceOnly* attack performs much better when the same audio content is present in the training and test data. However, such a restriction on content cannot be placed on the audio content hence we modify the audio samples to contain commonly used syllables in the English language. To study this approach we carry out an investigation using 345 data samples comprising of single-

syllable audio data collected in 3 different locations (i.e., two indoor bedrooms and one outdoor location-terrace). We noted that $\mathit{ForYourVoiceOnly}$ reached an average accuracy of 99% in predicting between the outdoor location and any one of the indoor locations. Further, when trying to classify between all the three locations our attack resulted in an accuracy of 95.2%. This experiment helped confirm that $\mathit{ForYourVoiceOnly}$ can train on messages of hardly a second and shows that 10 seconds worth of data at each location position is also more than enough to perform location classification.

8.3.4. EXTRACTED WORDS FROM VOICE MESSAGES

In a real scenario, voice messages can be of any length. To assess that our approach can be applied to a real-world context, we carried out a preliminary evaluation on 345 audio samples of words extracted from complex voice messages in the *Complete Profiling* scenario. Also, we reduced the number of rooms in our pool size to 3 (i.e., two indoor bedrooms and one outdoor location-terrace). We noted that $\mathit{ForYourVoiceOnly}$ reached an average accuracy of 99% in predicting between the outdoor location and any one of the indoor locations with both datasets, one containing mono-syllable audio messages and the other with extracted syllables. Further, when trying to classify between all the three locations our attack resulted in an accuracy of 94% with extracted syllables and 95.2% with monosyllabic audio messages. These results demonstrate that $\mathit{ForYourVoiceOnly}$ can be applied in real-world contexts by extracting single words from a complex voice message.

EXTRACTING WORDS FROM COMPLEX VOICE MESSAGES

For our above experiment, we wanted to compare the behavior of $\mathit{ForYourVoiceOnly}$ when the data used are monosyllable messages vs syllables extracted from complex voice messages. To extract these specific syllables from the audio message we used a manual method in combination with a speech to text service for which we utilized IBM Watson [1]. The service helped us obtain the timestamps which indicated when each word begins and ends in the audio stream. We then used these timestamps in conjunction with the ***ffmpeg*** command to extract the syllables and store them as separate audio files.

8.4. SPEAKER DATA

So far our experimental data were audio recordings recorded by people on their smartphone devices. This setup requires an attacker to work in conjunction with other attackers to help obtain data to train our model for the location profiling scenario and we wanted to avoid this dependency ideally. Given that this research was being carried out during a global pandemic further motivated us to seek a means of completely avoiding human participants in our experiment for all scenarios. Hence, we attempted to use speakers in place of human participants and collected data to perform some preliminary experiments. These speakers would play audio recordings of humans collected from locations unknown to us. The audio samples are WhatsApp audio messages, in this setup, we play the previously recorded messages on our speaker and recorded the message with a smartphone device keeping the distance between them nearly 15 cm.

[1] https://cloud.ibm.com/docs/speech-to-text/getting-started.html

8.4.1. PHONE SPEAKER

In this setup we used the easiest available speaker at hand - the phone speaker to perform data collection. We carried out a preliminary investigation on 300 audio samples collected at 2 different locations (i.e., two bedrooms). For collecting data a phone speaker was used and this phone was held horizontally with the phone speaker facing the screen of the device recording the audio message. We noted that $\mathit{ForYourVoiceOnly}$ reached an average accuracy of 58% in predicting the class between the two locations. These results demonstrate that $\mathit{ForYourVoiceOnly}$ performs poorly when the models are trained with phone speaker data. This is expected due to the difference in the way sound is produced by humans vs the way sound is generated in devices. Moreover, the phone speaker may not be very powerful and has limited capabilities depending on the model.

8.4.2. JBL GO SPEAKER

In this setup, we performed a very similar experiment as explained in 8.4.1. The only difference is that in this experiment the speaker used is a wireless Bluetooth speaker - JBL GO. Here the volunteer holds the device connected via Bluetooth to another device containing the recording. Then the volunteer plays the recording on the JBL speaker and records the audio with a smartphone device held at nearly 15 cm distance from the speaker. This setup is to recreate as closely as possible the scenario wherein we used humans for recording. We further perform two types of experiments discussed in the following subsections. From these subsections, we see that the results are not very conclusive and the data also is not good enough to give us stable values which will help us make incontestable observations. Hence, we went ahead and used human participants in our final data collection.

RECORDING AND TESTING ON SAME DEVICE

In this scenario, the speaker plays the audio recorded on the same device in which the test audio is recorded. We collected a total of 500 audio data samples with the help of two volunteers at two different locations (i.e., two indoor rooms). We used the following devices - OnePlus 5T and OnePlus 3. We noted that $\mathit{ForYourVoiceOnly}$ reached an average accuracy of 75% in predicting the class between the two locations.

Testing on an unknown speaker

Since $\mathit{ForYourVoiceOnly}$ is still able to retain some amount of distinguishing capability we tested the model by training and testing on speakers who are not present in training data. For this experiment, we collected 200 datapoints from the same two locations as mentioned before. This results in $\mathit{ForYourVoiceOnly}$ having an accuracy of 87%.

RECORDING AND TESTING ON DIFFERENT DEVICE

In this scenario, the speaker plays the audio recorded on a device as which is different from the one the test audio is recorded on. We obtained a total of 400 audio samples collected with the help of two volunteers at two different indoor locations (i.e., two bedrooms). For this setup, we achieved an accuracy of nearly 66% in distinguishing between

the two rooms.

Testing on an unknown speaker
Here we test the above model setup by testing on speakers who are not present in training data. For this setup, we obtain an accuracy of 65%.

9

EXPERIMENTAL SETTINGS

In this section, we provide details about the procedure followed during data collection along with the characteristics of the obtained dataset. We further provide a comprehensive overview of the machine learning models and parameters we used to demonstrate the efficacy of our proposed attack.

9.1. DATA COLLECTION

We performed our data collection at four different real locations. The layouts of these locations are depicted in the figures below. In particular, we considered three indoor locations I1 (Figure 9.1), I2 (Figure 9.2), and I3 (Figure 9.3), and one outdoor location O1 9.4. Since our goal is to recognize the specific location (or the specific position) from which a voice message is sent, for indoor locations we decided to consider the worst-case where the rooms have a similar layout and furnishings (i.e., bedrooms). Within each of the indoor locations, we further identify 5 different recording positions: south-east corner (P1), south-west corner (P2), north-west corner (P3), north-east corner (P4), and center (P5). While for O1, we identified a central recording position only (P5).

The data collection process involved 15 participants (5 males and 10 females aged 20 to 59 years). We ensured that the participants held their phones at a distance of about 15 cm from their face at chin level as shown in Figure 6.1. While recording, only the participant was present in the room and the room doors and windows were closed. To create a more realistic dataset we asked the participants to use their own smartphone devices. During the collection phase, the participants recorded 30 different voice messages using WhatsApp in all the locations at each position (positions are marked in the location layouts). This results in a total of 150 recordings per indoor location and 30 recordings for the outdoor location. We collected a total of 7200 WhatsApp voice messages, corresponding to 480 recordings per participant.

All the recorded WhatsApp voice messages have a one-second duration (i.e., the minimum duration of a WhatsApp voice message) and contain a single word. Specifically, for each position the participants recorded 30 voice messages: 10 pronouncing the word *and*, 10 pronouncing the word *of*, and 10 pronouncing the word *the*. We selected these words based on the OEC and COCA ranks for most commonly used words during an English conversation [15, 17]. We divided the 30 recordings at a single position into three sequences of 9-12-9. The participant starts the data collection from position P1, recording 9 voice messages at P1 (i.e., 3 voice messages per word). Once concluded with this step, the participant moves to P2 in the same location and records 9 voice messages again. After all the five positions are covered in sequence, the participant starts the procedure again from P1 recording 12 voice messages (i.e., four voice messages per word). Finally, the participant concludes the data collection with a final set of 9 voice messages per position before moving to the next location. For the O1 location, the participant recorded 30 voice messages from the same position (i.e., P5) but in a sequence of 9-12-9 in a discontinuous manner.

Finally, along with these recordings information about the room such as its dimensions, the finish, the tiling, etc. is noted. This information depicts the features that could possibly affect recordings taken in that room. Also for each recording session apart from the room we also maintain details regarding the participant such as the gender and also the phone model used during the recording process.

9.2. MACHINE LEARNING MODELS

To identify the location and the specific position in a location of a voice message, we tested four multi-class classifiers: Linear Discriminant Analysis (LDA), Logistic Regres-

sion (LR), Ridge Classifier (RC), and Support Vector Machine (SVM). Based on the attack scenario we applied different strategies to split the data into training, validation, and testing sets:

- **Complete Profiling**: To evaluate the performance of our approach, we apply (for each participant) a nested-cross fold validation. In the outer loop, we use a stratified 5-fold cross-validation on the 480 voice messages recorded by the participant, resulting in 384 recordings in training and 96 in testing per fold. In the inner loop, we apply a stratified 3-fold cross-validation on the 384 training recordings, obtaining 256 recordings in training and 128 recordings in validation per fold.

- **Location Profiling**: For this experiment we consider the entire dataset comprising of 7200 audio recordings, and we apply a nested cross-fold validation. For the outer loop, we apply a user-independent leave-one-out cross-validation, obtaining a testing set containing the recordings of a single participant (i.e., 480). Similarly in the inner loop, we apply a user-independent leave-one-out cross-validation on the other 14 participants, obtaining a training set of 13 participants (i.e., 6240 recordings) and a validation set of one participant (i.e., 480 recordings) for each iteration.

- **User Profiling**: In this scenario, we consider the dataset of each participant individually, as for the*Complete Profiling scenario*. Also here we apply a nested-cross fold validation, but differently to the *Complete Profiling scenario*, we use a group-k-fold to split the dataset into subsets based on the recording location. We use a group 5-fold cross-validation in the outer loop and a group 4-fold cross-validation for the inner loop. In this way, we split data recorded within the same room into subsets corresponding to each of the 5 recording positions (i.e., P1, P2, P3, P4, and P5). Using this configuration both the validation and the test sets consist of one subset each, while the training set contains the remaining positions. The recordings from location O1 are excluded from this scenario since they all come from the same location position (i.e., P5).

We explored different hyper-parameters by using grid search on all the considered classifiers. In particular, for LDA we vary the solver over [*svd*, *lsqr*, *eigen*]. For LR we vary the solver in [*newton-cg*, *lbfgs*, *liblinear*] and the C value in the range [10^{-3}, 10^{-2}, ..., 10^{1}]. For RC we vary α from 0.1 to 0.9 with a step size of 0.1, and from 1 to 10 with a step size of 1. Finally, for SVM we tune the values parameter C in the range [10^{-1}, 10^{0}, ..., 10^{3}], and γ in the range [10^{-4}, 10^{-3}, ..., 10^{0}].

9.3. CROSS VALIDATION

Since we have a very limited amount of data available with us we resample the data to help ensure that the model performs well on data not known to it during training. We have used nested cross-validation to achieve this. We chose to use this technique as we wanted to optimize the hyperparameter values of the models alongside performing validation. Hyperparameter tuning can overfit the given data and result in an idealistic assessment of a model that should not be utilized to select our model. Nested CV helps

to ensure that we do not overfit our data by tuning the parameters on the same dataset which is then used to evaluate the performance of the tuned model. In this technique, hyperparameter tuning is performed on a model and the evaluation is nested within a more wide-ranged outer k-fold cross-validation procedure as shown in Figure 9.6. While the technique remains constant for all scenarios the realization varies slightly as mentioned in the previous section.

For soft and hard voting classifier [1], we used a faster nested cross-validation method as shown in Figure 9.5. Here, we use the inner loop data to train our voting classifier model and obtain the best-tuned classifier for the given data from the ensemble list. This classifier which has tuned values for its hyperparameters is then used to train the training data of the outer loop and tested on the test data of the outer cross-validation loop.

[1] These classifiers were only used in the preliminary experiments. This was because these classifiers consumed a lot of time in comparison to the provided increase in precision of $\mathit{ForYourVoiceOnly}$ accuracy.

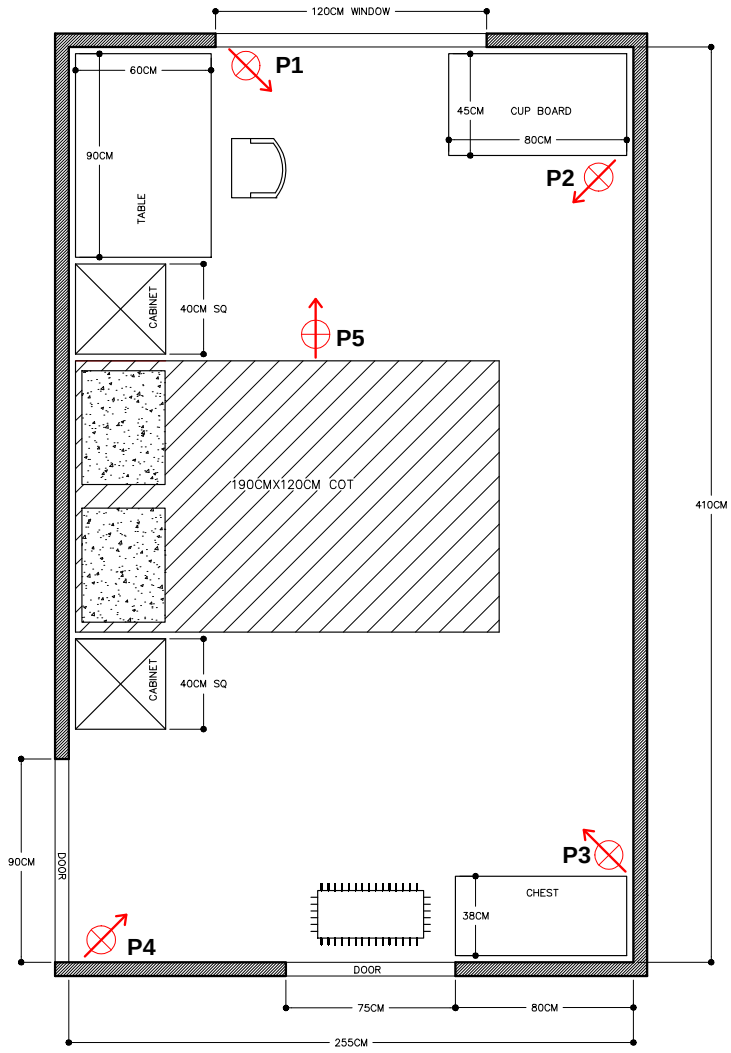


Figure 9.1: Location layout and recording positions with orientation considered in the data collection in Indoor Location I1.

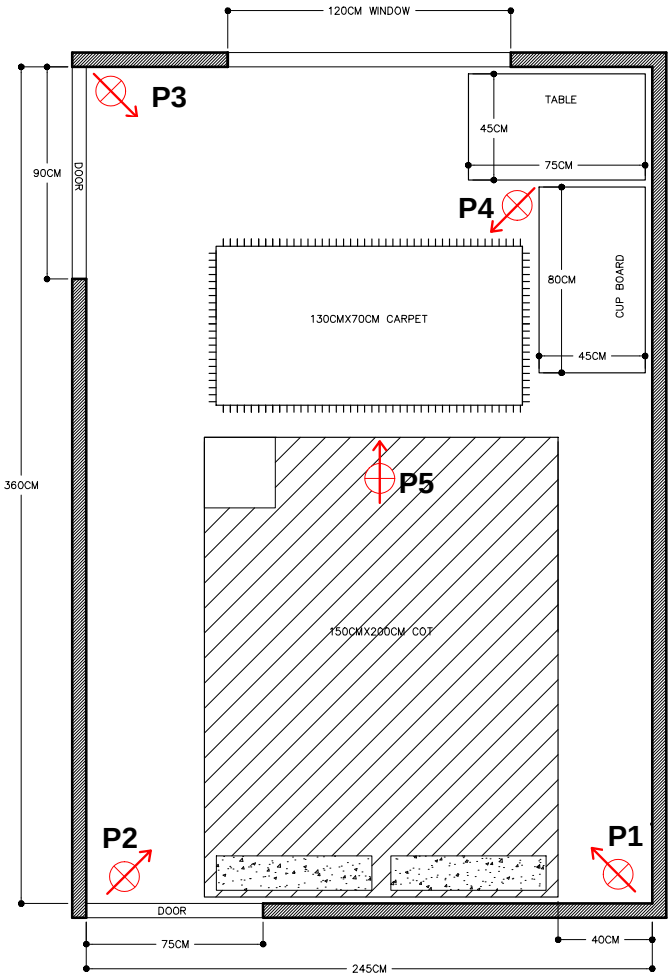


Figure 9.2: Location layout and recording positions with orientation considered in the data collection in Indoor Location I2.

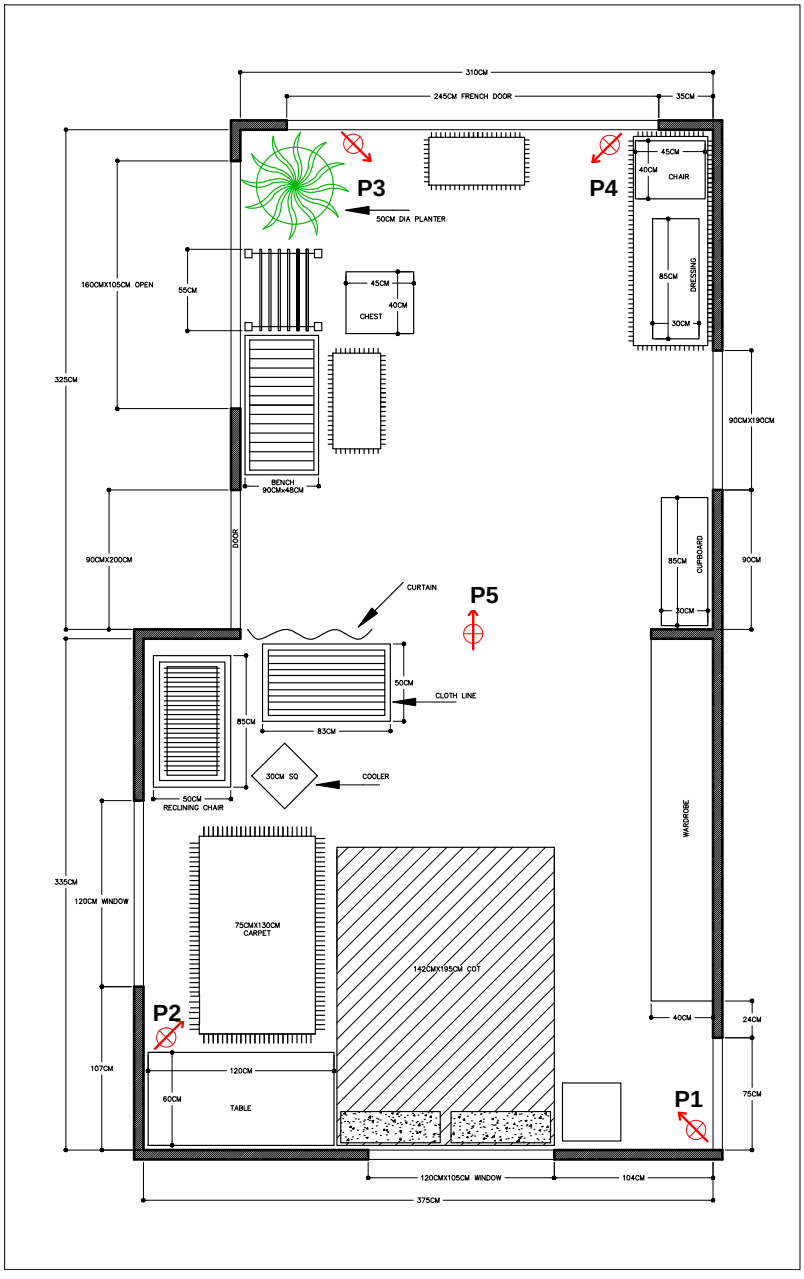


Figure 9.3: Location layout and recording positions with orientation considered in the data collection in Indoor Location I3.

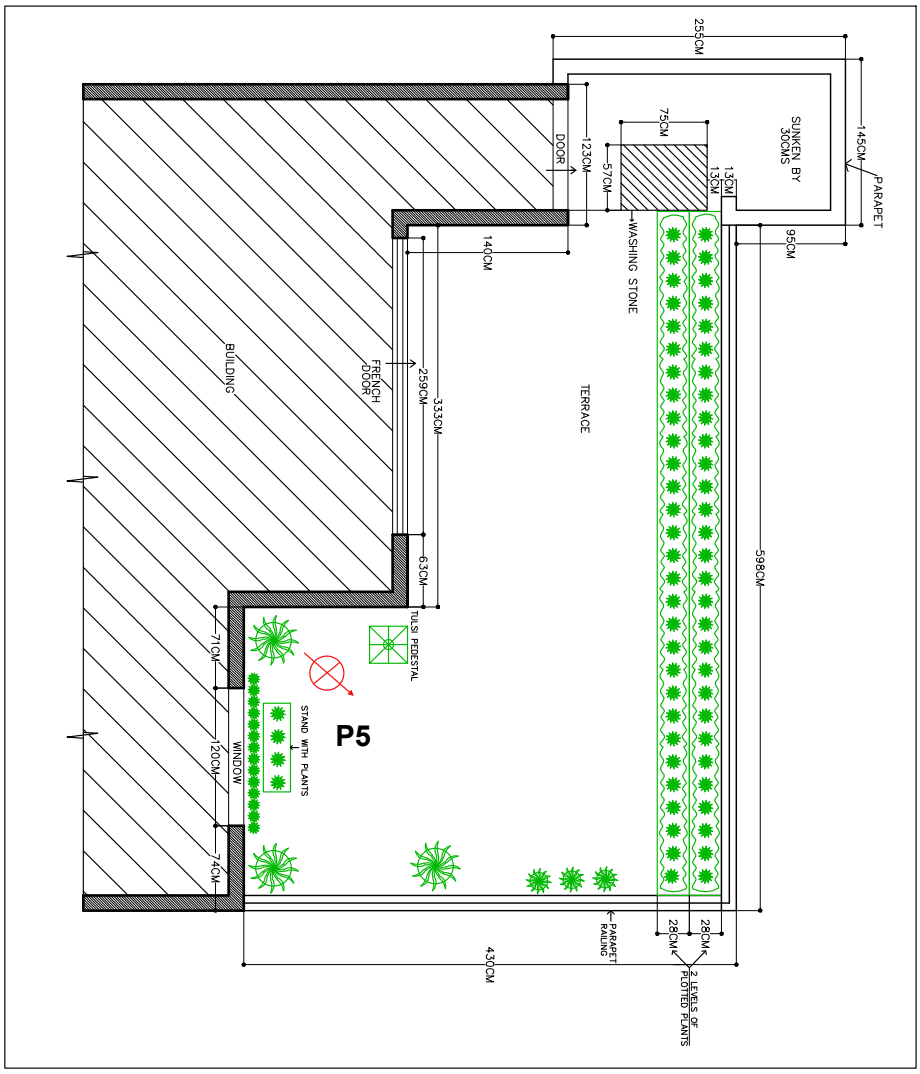


Figure 9.4: Location layout and recording positions with orientation considered in the data collection in Outdoor Location O1.

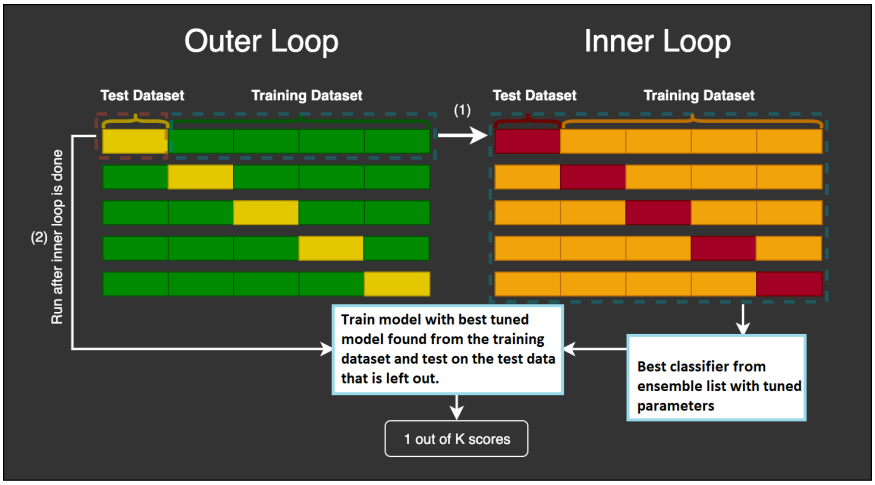


Figure 9.5: Nested Cross Validation for Voting Classifiers

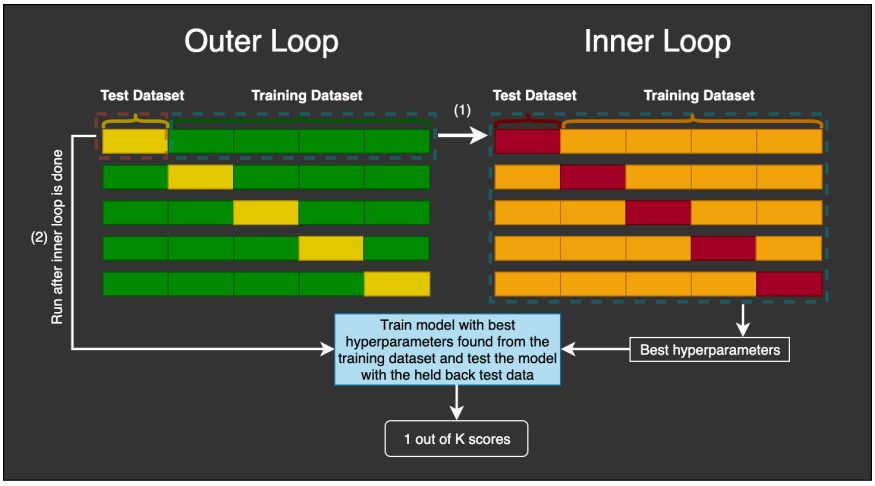


Figure 9.6: Nested Cross Validation

10

Experimental Results

In this section we report and discuss the results achieved by $\mathcal{F}or Your Voice Only$ *in the three attack scenario based on the attack goal: location 10.1 or position 10.2. Finally, in Section 10.3 we introduce our survey set up, and then we compare the accuracy of* $\mathcal{F}or Your Voice Only$ *vs. that of humans. The detailed results can be found in the Appendix A.*

10.1. LOCATION INFERENCE

In Table 10.0 we show the performance of the classifiers in identifying the location according to the attack scenario, considering the worst case for each scenario (i.e., 4 locations for the *Complete Profiling* and *Location Profiling* scenarios, and 3 locations for the *User Profiling* scenario).

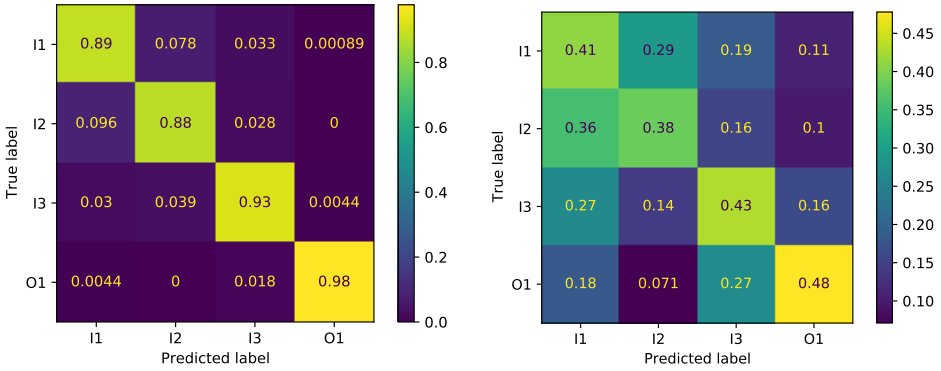
Table 10.0: Average accuracy of $\mathcal{F}orYourVoiceOnly$ attack for location inference in different attack scenarios.

Scenario	LDA	LR	RC	SVM
Complete Profiling	0.85 (0.06)	0.85 (0.06)	0.83 (0.06)	0.87 (0.05)
Location Profiling	0.41 (0.11)	0.39 (0.10)	0.43 (0.09)	0.35 (0.00)
User Profiling	0.80 (0.09)	0.33 (0.04)	0.32 (0.03)	0.33 (0.03)

The scenario where the classifiers perform best is the *Complete Profiling* scenario, where the attacker has the maximum information available. The results show that in this scenario all classifiers have accuracy higher than 83%. In particular, the SVM manages to reach an accuracy of 87%. On the contrary, in the *Location Profiling* scenario, there is a consistent drop in performance. In this case, the best classifier is the RC, which reaches an accuracy of 43% (i.e., 18% above the chance level). Lower performance can be attributed to multiple factors:

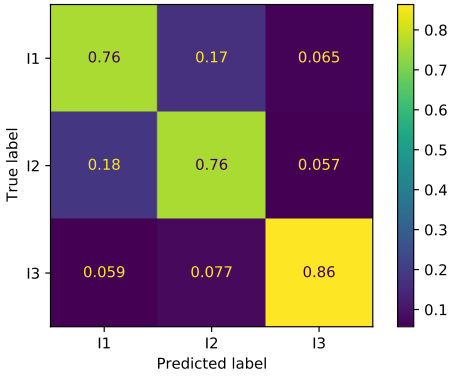
- Device: The participants used different phones during data collection, the absence of the model in the training set may be a contributing factor to a reduced accuracy on new test data.
- Training Size: The number of users in training is not enough to ensure sufficient variability in the training features.
- Voice Uniqueness: The distinctiveness of the victim's vocal characteristics cannot be completely replaced, and their absence in training is reflected in the testing performance.
- Variable Background Noise: The users recorded on different days at different times over a period of 1 month. This impacts the background noise present during recording which may also lead to a decreased accuracy.

The importance of the victim's voice for the attacker is supported by the results obtained for the *User Profiling* scenario, where the attacker has voice messages from the victim but does not know the specific recording location. In this case, LDA achieves an accuracy of 80% (i.e. only 7% less than in the Complete Profiling scenario). It is interesting to note that in the User Profiling scenario, unlike the others, there is a classifier that outperforms the others. In Figure 10.1 we show the confusion matrices of the best model per scenario in the location classification. It is interesting to note that in all three attack scenarios the locations I1 and I2 are confused with each other. This is due to the similar layout of the two locations (see Figures 9.1 and 9.2). The background noise is instead discriminant for the identification of the external location (i.e., O1). O1 is generally classified better, reaching an accuracy up to 98% in the *Complete Profiling* scenario.



(a) *Complete Profiling scenario.*

(b) *Location Profiling scenario*



(c) *User Profiling scenario*

Figure 10.1: $\mathcal{F}orYourVoiceOnly$ confusion matrices for the best models.

Further, we analyzed the influence of the number of locations of interest (i.e., the number of classes to be predicted) on the accuracy of the classification. In the *Complete Profiling* scenario we obtain an average accuracy of 99% when we classify an audio message between the outdoor location O1 and one of the indoor locations (i.e. I1, I2, and I3). While when we classify messages between two indoor rooms we achieve an accuracy ranging from 89% to 95% on this task. Also in *Location Profiling* scenario, we obtain a higher accuracy if we reduce the location of interest considering O1 and an indoor location. In this case, $\mathcal{F}orYourVoiceOnly$ correctly predicts the location with an average accuracy of 80%. While for the prediction of internal location pairs the accuracy remains rather low, ranging from 57% between I1 and I2 to 66% between I1 and I3. Finally, considering the *User Profiling* scenario, reducing the locations of interest to two leads to an average accuracy of 87% in predicting the correct recording location.

Finally, we evaluated $\mathcal{F}orYourVoiceOnly$ by training the models on a single word, splitting the dataset into three subsets of 2400 audio recordings each containing the

words "*and*", "*of*" and "*the*". Figure 10.2 depicts the variation of the accuracy of our attack in the *Complete Profiling* scenario between all the locations I1, I2, I3, and O1 using different classifiers and different words. Results show that there are no significant differences between models trained on the specific word and those trained on all words (i.e., combined).

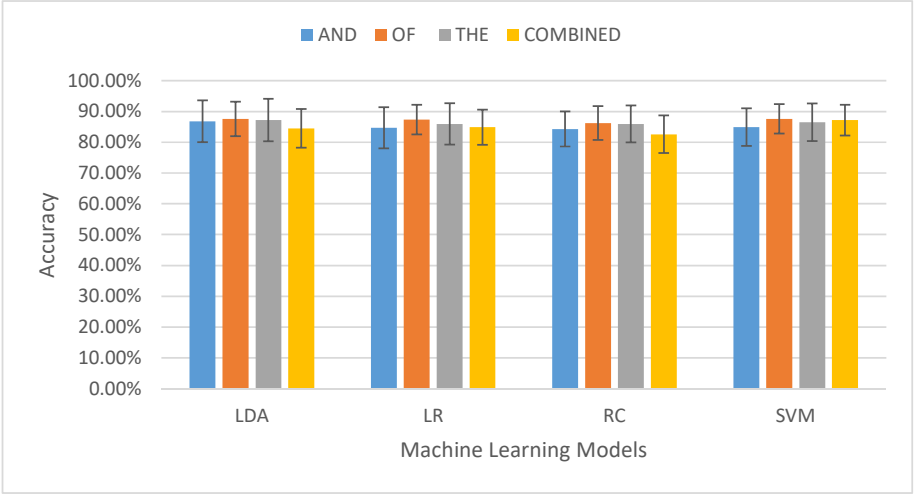


Figure 10.2: Performance of machine learning models in classifying the four locations in *Complete Profiling* scenario when trained specifically with one word and all the words (i.e., combined).

10.2. POSITION INFERENCE

In Table 10.2 we show the performance of the classifiers in identifying the specific position according to the attack scenario, considering the worst case (i.e., 16 positions - five for each indoor location and one for the outdoor location). Unlike *Location Inference*, here we consider only two attack scenarios (i.e., *Complete Profiling* and *Location Profiling*), since the *User Profiling* scenario assumes that the attacker has no information about the specific position in training. As in Location Inference, even for the position inference, the scenario where the classifiers perform best is the *Complete Profiling* and SVM resulted in the best classifier scenario with an accuracy of 61%. Contrarily, in *Location Profiling* scenario models performance is slightly above chance (i.e., 0.0625). The increase in the number of classes to be predicted and the factors already highlighted in Section 10.1 (i.e., device, training size, and voice uniqueness) further amplify the performance drop.

In Figure 10.3 we show the confusion matrix of the best model in the *Complete Profiling* scenario (i.e, SVM).

As expected, the model manages to accurately predict O1 (i.e., 98%), demonstrating that this is a trivial task for our attack in this scenario. Regarding the internal locations Figure 10.3 shows a concentration of classification errors in the positions belonging to the true location. In particular, the classification of I3 positions shows less accuracy than I1 and I2. We believe that this can be traced back to the layout of the room. In fact, I3

Table 10.2: Average accuracy of $\mathcal{F}orYourVoiceOnly$ attack for position inference in different attack scenarios.

Scenario	LDA	LR	RC	SVM
Complete Profiling	0.57 (0.09)	0.55 (0.09)	0.49 (0.08)	0.61 (0.09)
Location Profiling	0.13 (0.04)	0.13 (0.04)	0.13 (0.04)	0.07 (0.00)

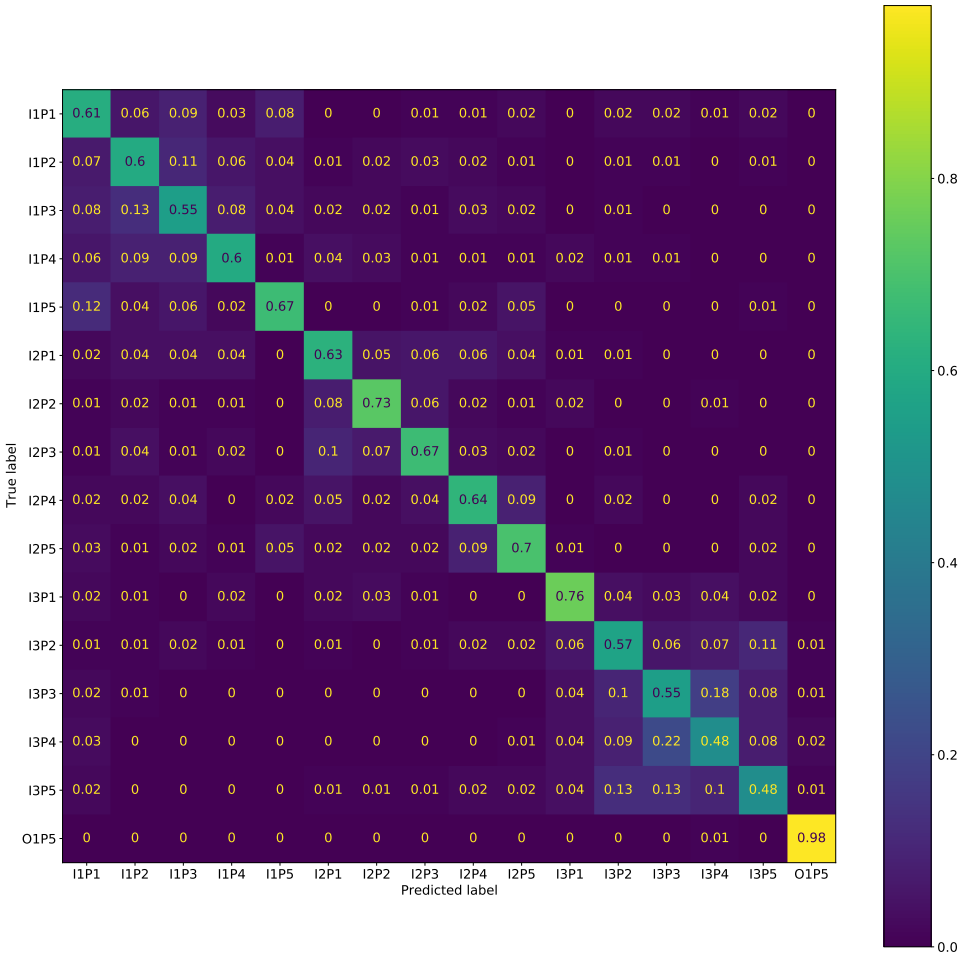


Figure 10.3: Confusion matrix for specific position inference with for I1, I2, I3 and O1 locations in *Complete Profiling* scenario.

has more than twice the surface area of I1 and I2, and the spaces between the recording points and the walls or furniture are much wider. This could lead to a reduction in reverberation and therefore make the recordings more similar. In addition, the best performing position in I3 is P1, which is the recording position with the least open field compared

to the other four positions. I1 and I2 generally present better results, but again we can see how room size affects the prediction of the specific location. I2 measures about 2 square meters less than I1 and has a 7% higher average accuracy.

10.3. SURVEY

In addition to studying how well our $\mathscr{F}orYourVoiceOnly$ model works with the collected audio data we thought it would be of value to study how hard the task would be if we had to solely rely on the human auditory system. For this purpose, we formulated a survey for volunteers to fill out.

10.3.1. SURVEY ORGANIZATION

To simulate the three scenarios discussed previously we divided our survey into two parts. The first survey contained 96 audio recordings pertaining to 6 speakers with samples associated with the 4 different locations [1] to train the participants. This survey was to simulate the *Complete Profiling* and *User Profiling* scenarios. Once the training is completed each participant was asked to classify 24+18 test audio samples via the survey [2]. The first 24 samples were audio recordings of the same 6 speakers in rooms and positions present in the training. The latter 18 test samples belonged to the 6 speakers in training recorded at the same locations but from new positions within these locations. To simulate the *Location Profiling* scenario we trained the participants with a new dataset comprising of 72 audio recordings pertaining to 6 (different from the six speakers used in the *Complete/User Profiling* setup) speakers with samples pertaining to the 4 different locations [3]. Once the participants listen to the audio recordings in training they were requested to fill out a second survey [4] consisting of 12 test audio samples. These audio samples belonged to 3 (these 3 speakers are neither present in the training nor were they part of the dataset used for the *Complete/User Profiling* setup) new speakers in the locations and positions covered in the training data. Here, the participants were expected to fill out the survey classifying test data after they listened to the training data which was made available to them via a website as mentioned. However, the participants could have this website open while filling out the survey for additional support. In total we had 29 participants fill out the first survey regarding *Complete/User Profiling* the gender distribution is displayed in 10.4 and age distribution of the participants are shown in Figure 10.5. For the second survey corresponding to *Location Profiling* we had a total of 21 participants and their gender/age distribution is shown below in 10.6 and 10.7 respectively.

10

10.3.2. SURVEY RESULTS

Here we report the results we obtained in our survey. We further draw a comparison between the results obtained by $\mathscr{F}orYourVoiceOnly$ vs. what we achieved using just human participants.

[1] https://mailarpitar.wixsite.com/location-guessing/dataset1
[2] https://forms.gle/NjBH2EAV7TjpyUit9
[3] https://mailarpitar.wixsite.com/location-guessing/dataset2
[4] https://forms.gle/3rk5wtwjnMVc6Hc36

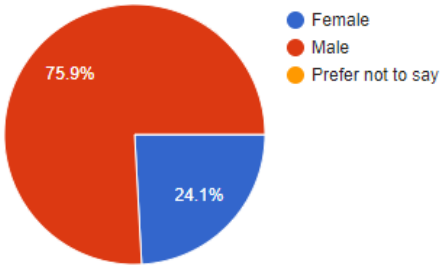


Figure 10.4: Gender distribution of participants of the *Complete/User Profiling* survey

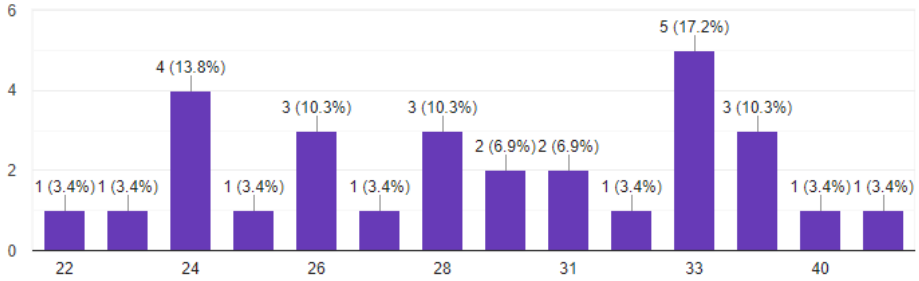


Figure 10.5: Age distribution of participants of the *Complete/User Profiling* survey

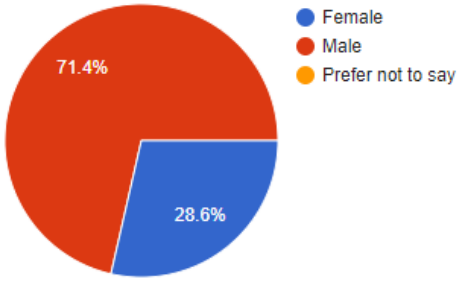


Figure 10.6: Gender distribution of participants of the *Location Profiling* survey

- Complete Profiling: The survey had 29 participants each classifying 24 test audio samples. This resulted in an accuracy of nearly 24% (The probability of guessing correctly is 25%). To see how well the participants can perform indoor-outdoor classification we tested how well the participants could classify audio recordings belonging to the location O1 - the achieved accuracy is only 60%. In 10.8 we show

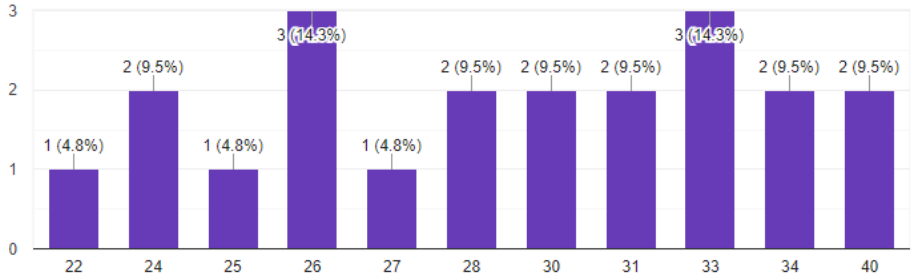


Figure 10.7: Age distribution of participants of the *Location Profiling* survey

how poorly humans perform when compared to $\mathcal{F}orYourVoiceOnly$.

- Location Profiling: The survey had 21 participants each classifying 12 test audio samples. Evaluating this survey we obtain an accuracy of 35% for the task (The probability of guessing correctly is 33.33%). Given the test sample of an outdoor location, the participants achieved an accuracy of nearly 78% in correctly identifying the recording location as outdoors. Figure 10.9 depicts how well machine learning classifiers can discern audio recording locations in comparison to humans.
- User Profiling: The survey had 29 participants each classifying 18 test audio samples. This setup resulted in an accuracy of approximately 25% (The probability of guessing correctly is 25%). To see how well humans can classify indoor and outdoor audio recording we only had test cases pertaining to the indoors however we still had 17.4% misclassifying the test sample. Figure 10.10 depicts how much better an ML classifier can utilize reverberations and ambient noise to correctly classify the location in which audio was recorded.

These results demonstrate to us how difficult it is to leverage the leaked data in audio messages using solely human hearing abilities. Also, the indoor-outdoor classification is relatively simpler due to the presence of distinctive background noises. This implies that this task is relatively simple even for the human participants and this is confirmed by the accuracy we obtained for this task in comparison to the other tasks.

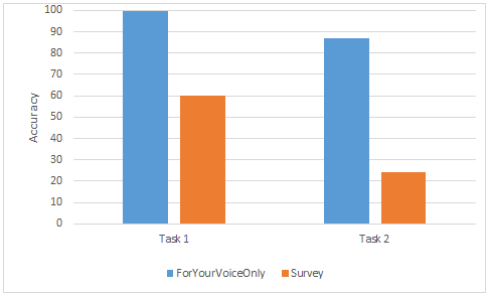


Figure 10.8: Comparison of task accuracy between humans and *ForYourVoiceOnly* for *Complete Profiling*. Here task 1 is indoor-outdoor classification and task 2 corresponds to the room classification between all 4 known locations

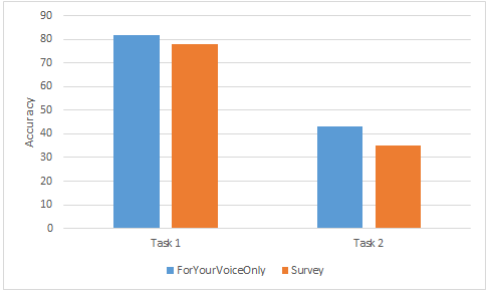


Figure 10.9: Comparison of task accuracy between humans and *ForYourVoiceOnly* for *Location Profiling*. Here task 1 is indoor-outdoor classification and task 2 corresponds to the room classification between all 4 known locations.

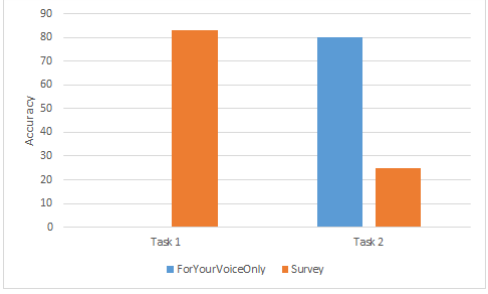


Figure 10.10: Comparison of task accuracy between humans and *ForYourVoiceOnly* for *User Profiling*. Here task 1 is indoor-outdoor classification and task 2 corresponds to the room classification between all 3 indoor locations.

11

CONCLUSION

In this report, we proposed $\mathit{ForYourVoiceOnly}$, a new attack on voice messages to infer the recording location. $\mathit{ForYourVoiceOnly}$ leverages attributes such as reverberation and ambient noises which inadvertently get recorded along with audio messages. We showed the effectiveness of our attack in three realistic attack scenarios: (i) the attacker has previous recordings of the victim in all the selected locations (ii) the attacker has no previous recording of the victim's voice messages (iii) the attacker has previous voice messages of the victim knowing the location they were recorded but does not know the specific position. We demonstrated our attack considering 7,200 voice messages from 15 different users and four environments (i.e., three bedrooms and a terrace). We showed how the possession of audio messages from the victim in known locations greatly increases the performance of our attack. $\mathit{ForYourVoiceOnly}$ can infer the location of the user among a pool of four known environments with up to 85% accuracy. Moreover, our approach reaches an average accuracy of 93% in discerning between two rooms of similar size and furniture (i.e., two bedrooms), and an accuracy of up to 99% in classifying indoor and outdoor environments.

10.1. LIMITATIONS

We believe that the proposed work can be a starting point for developing environment recognition from voice messages and can overcome the limitations of $\mathit{ForYourVoiceOnly}$. First, the collection of new datasets would allow for more consolidated results and the application of more powerful feature extraction and prediction techniques (e.g., deep learning). It would be useful to have a more diverse dataset in terms of languages, gender, age, and nationality. Further, we only used a single voice messaging application to collect data, introducing more such applications or the use of audio call recordings in the dataset would also provide more useful insight.

The collection of new datasets would also be beneficial for assessing the effect of noisier environments. We made several restrictions during recording such as having no other member in the rooms during recording, the recordings were done in a relatively

quiet and less crowded location. Hence, we expect the behavior to be affected when the noise increases. This can be either detrimental or instrumental depending on whether valuable information is obscured or if the noise is indicative of that particular location.

10.2. FUTURE WORKS

Based on our research work, we believe that there are various directions in which potential research works can be carried out, such as developing effective countermeasures or applying our approach to other scenarios.

- We also think that combining the results of multiple test audio samples will help in improving the results obtained by $\mathit{ForYourVoiceOnly}$. The classification outputs may be combined based on the probability assigned by the classifier.
- It would also be interesting to note whether changes in room furniture during the course of recording change the accuracy with which the model identifies the recording location.
- Another intriguing aspect to analyze is whether changes to voice made with voice modification tools affect our proposed system. Or whether such tools can be used as a countermeasure to our proposed attack.
- We also see many possible directions in which countermeasures to our proposed attack may be developed. Some of these are -
  - The use of noise to obscure the leaked information in the audio messages. The noise may also be applied selectively to higher and lower frequencies outside the hearing range so as not to impact the quality of the voice message. This method may prove to act as a countermeasure as we noted variations in the audio signals in the ultrasonic and infrasonic ranges at different locations and positions.
  - Another straight forward measure may be shielding of the microphone during recording.
  - It may also be rather interesting to note if the audio can be filtered so as to retain only the primary sound source and eliminate all other signals from the recorded audio.

BIBLIOGRAPHY

[1] Dominic W Massaro. "Preperceptual images, processing time, and perceptual units in auditory perception." In: *Psychological review* 79.2 (1972), p. 124.

[2] Lawrence Rabiner. "Fundamentals of speech recognition". In: *Fundamentals of speech recognition* (1993).

[3] Huan Liu and Rudy Setiono. "Chi2: Feature selection and discretization of numeric attributes". In: *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*. IEEE. 1995, pp. 388–391.

[4] Brian D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996, 91–120.

[5] Eric Scheirer and Malcolm Slaney. "Construction and evaluation of a robust multifeature speech/music discriminator". In: *1997 IEEE international conference on acoustics, speech, and signal processing*. Vol. 2. IEEE. 1997, pp. 1331–1334.

[6] Mario Köppen. "The curse of dimensionality". In: *5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*. Vol. 1. 2000, pp. 4–8.

[7] Claude Chibelushi, Farzin Deravi, and John Mason. "A review of speech-based bimodal recognition". In: *Multimedia, IEEE Transactions on* 4 (Apr. 2002), pp. 23 –37. DOI: 10.1109/6046.985551.

[8] Vesa Peltonen et al. "Computational auditory scene recognition". In: *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 2. IEEE. 2002, pp. II–1941.

[9] Michael Cowling and Renate Sitte. "Comparison of techniques for environmental sound recognition". In: *Pattern recognition letters* 24.15 (2003), pp. 2895–2907.

[10] Guodong Guo and Stan Z Li. "Content-based audio classification and retrieval by support vector machines". In: *IEEE transactions on Neural Networks* 14.1 (2003), pp. 209–215.

[11] Hyoung-Gook Kim, Nicolas Moreau, and Thomas Sikora. "Audio classification based on MPEG-7 spectral basis representations". In: *IEEE Transactions on Circuits and Systems for Video Technology* 14.5 (2004), pp. 716–725.

[12] Lei Chen, Sule Gunduz, and M Tamer Ozsu. "Mixed type audio classification with support vector machine". In: *2006 IEEE International Conference on Multimedia and Expo*. IEEE. 2006, pp. 781–784.

[13] Selina Chu, Shrikanth Narayanan, and C-C Jay Kuo. "Environmental sound recognition with time–frequency audio features". In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.6 (2009), pp. 1142–1158.

[14] Michael Backes et al. “Acoustic Side-Channel Attacks on Printers.” In: *USENIX Security symposium*. Vol. 9. 2010, pp. 307–322.

[15] Mark Davies. “The Corpus of Contemporary American English as the first reliable monitor corpus of English”. In: *Literary and linguistic computing* 25.4 (2010), pp. 447–464.

[16] Dalibor Mitrović, Matthias Zeppelzauer, and Christian Breiteneder. “Features for content-based audio retrieval”. In: *Advances in computers*. Vol. 78. Elsevier, 2010, pp. 71–150.

[17] Angus Stevenson. *Oxford dictionary of English*. Oxford University Press, USA, 2010.

[18] Junzhao Du et al. “Catch you as i can: indoor localization via ambient sound signature and human behavior”. In: *International Journal of Distributed Sensor Networks* 9.11 (2013), p. 434301.

[19] Çigdem Okuyucu, Mustafa Sert, and Adnan Yazici. “Audio feature and classifier analysis for efficient recognition of environmental sounds”. In: *2013 IEEE International Symposium on Multimedia*. IEEE. 2013, pp. 125–132.

[20] Juan Rubén Delgado-Contreras et al. “Feature selection for place classification through environmental sounds”. In: *Procedia Computer Science* 37 (2014), pp. 40–47.

[21] “Chapter 4 - Audio Features”. In: *Introduction to Audio Analysis*. Ed. by Theodoros Giannakopoulos and Aggelos Pikrakis. Oxford: Academic Press, 2014, pp. 59–103. ISBN: 978-0-08-099388-1.

[22] Yohan Petetin, Cyrille Laroche, and Aurélien Mayoue. “Deep neural networks for audio scene recognition”. In: *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE. 2015, pp. 125–129.

[23] Daniel Walnycky et al. “Network and device forensic analysis of android social-messaging applications”. In: *Digital Investigation* 14 (2015), S77–S84.

[24] Elsa Ferreira Gomes, Fábio Batista, and Alípio M Jorge. “Using smartphones to classify urban sounds”. In: *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering*. 2016, pp. 67–72.

[25] Huy Phan et al. “Learning representations for nonspeech audio events through their similarities to speech patterns”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.4 (2016), pp. 807–822.

[26] James Traer and Josh H McDermott. “Statistics of natural reverberation enable perceptual separation of sound and space”. In: *Proceedings of the National Academy of Sciences* 113.48 (2016), E7856–E7865.

[27] Hamid Eghbal-zadeh et al. “A hybrid approach with multi-channel i-vectors and convolutional neural networks for acoustic scene classification”. In: *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE. 2017, pp. 2749–2753.

[28] Banriskhem K Khonglah, KT Deepak, and SR Mahadeva Prasanna. “Indoor/Outdoor Audio Classification Using Foreground Speech Segmentation.” In: *INTERSPEECH*. 2017, pp. 464–468.

[29] Yusep Rosmansyah et al. “The microphone array sensor attack on keyboard acoustic emanations: Side-channel attack”. In: *2017 International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE. 2017, pp. 261–266.

[30] Noor Almaadeed et al. “Automatic detection and classification of audio events for road surveillance applications”. In: *Sensors* 18.6 (2018), p. 1858.

[31] Sergio Oramas et al. “Multimodal deep learning for music genre classification”. In: *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.* (2018).

[32] Wayne Xiong et al. “The Microsoft 2017 conversational speech recognition system”. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 5934–5938.

[33] S Chandrakala and SL Jayalakshmi. “Environmental audio scene and sound event recognition for autonomous surveillance: A survey and comparative studies”. In: *ACM Computing Surveys (CSUR)* 52.3 (2019), pp. 1–34.

[34] Inês Nolasco et al. “Audio-based identification of beehive states”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 8256–8260.

[35] Yusuf Ozkan and Buket D Barkana. “Forensic Audio Analysis and Event Recognition for Smart Surveillance Systems”. In: *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*. IEEE. 2019, pp. 1–6.

[36] Ivan Miguel Pires et al. “Recognition of Activities of Daily Living and Environments Using Acoustic Sensors Embedded on Mobile Devices”. In: *Electronics* 8.12 (2019), p. 1499.

[37] Ilia Shumailov et al. “Hearing your touch: A new acoustic side channel on smartphones”. In: *arXiv preprint arXiv:1903.11137* (2019).

[38] Federico Simonetta, Stavros Ntalampiras, and Federico Avanzini. “Multimodal music information processing and retrieval: survey and future challenges”. In: *2019 International Workshop on Multilayer Music Representation and Processing (MMRP)*. IEEE. 2019, pp. 10–18.

[39] Miad Faezipour and Abdelshakour Abuzneid. “Smartphone-based self-testing of covid-19 using breathing sounds”. In: *Telemedicine and e-Health* 26.10 (2020), pp. 1202–1205.

[40] Dias Issa, M Fatih Demirci, and Adnan Yazici. “Speech emotion recognition with deep convolutional neural networks”. In: *Biomedical Signal Processing and Control* 59 (2020), p. 101894.

[41] Zohaib Mushtaq and Shun-Feng Su. “Environmental sound classification using a regularized deep convolutional neural network with data augmentation”. In: *Applied Acoustics* 167 (2020), p. 107389.

[42] Jaime Ramírez and M Julia Flores. “Machine learning for music genre: multifaceted review and experimentation with audioset”. In: *Journal of Intelligent Information Systems* 55.3 (2020), pp. 469–499.

[43] Mishaim Malik et al. “Automatic speech recognition: a survey”. In: *Multimedia Tools and Applications* 80.6 (2021), pp. 9411–9457.

[44] URL: https://www.audiolabs-erlangen.de/resources/MIR/FMP/data/C1/FMP_C1_F23_FourInstruments.png.

[45] URL: https://superkogito.github.io/blog/SignalFraming.html.

[46] URL: https://ai.plainenglish.io/fischers-linear-discriminant-analysis-in-python-from-scratch-bbe480497504.

[47] URL: https://towardsdatascience.com/logistic-regression-explained-9ee73cede081.

[48] URL: https://stats.stackexchange.com/questions/402889/why-ridge-regression-only-decreases-slope-and-not-increases-it.

[49] URL: https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781838644338/5/ch05lvl1sec35/support-vector-machines.

A

APPENDIX

TASK	LDA	LR	RC	SVM
1	0.953333333	0.97	0.983333333	0.963333333
2	0.99	0.991	0.991133333	0.993266667
3	0.973333333	0.975533333	0.976533333	0.977666667
4	0.8682	0.847133333	0.843333333	0.849066667
5	0.876666667	0.859333333	0.868	0.885333333
6	0.93	0.918666667	0.932666667	0.928
7	0.951333333	0.931333333	0.950666667	0.936
8	0.865333333	0.849733333	0.858666666	0.858266667
9	0.987066667	0.989266667	0.989333333	0.995466667
10	0.648467	0.590867	0.5542	0.629067
11	0.6066	0.621733	0.649	0.622267
12	0.708	0.766667	0.753333	0.749333
13	0.558667	0.584	0.577333	0.597333
14	0.646	0.634	0.600667	0.64
15	0.662	0.634667	0.624	0.66
16	0.586667	0.574667	0.548	0.599333
17	0.618867	0.578667	0.545267	0.596867

Table A.1: The combined results of 15 participants for the Complete Profiling scenario with the AND syllable

TASK	LDA	LR	RC	SVM
1	0.95	0.983333	0.97	0.973333
2	0.991067	0.9944	0.9922	0.993267
3	0.989933	0.991067	0.9966	0.993267
4	0.875867	0.873733	0.862467	0.8762
5	0.868667	0.880667	0.880667	0.878667
6	0.944	0.939333	0.948667	0.930667
7	0.942667	0.941333	0.95	0.942
8	0.872933	0.8662	0.860933	0.8724
9	0.9906	0.987333	0.9906	0.9906
10	0.625067	0.611333	0.568333	0.6534
11	0.592	0.656	0.634667	0.670667
12	0.744	0.757333	0.768	0.742667
13	0.592	0.6	0.624	0.64
14	0.647333	0.646	0.608667	0.652
15	0.662667	0.668667	0.618	0.668
16	0.602	0.602	0.576667	0.639333
17	0.6108	0.603933	0.545267	0.6208

Table A.2: The combined results of 15 participants for the Complete Profiling scenario with the OF syllable

TASK	LDA	LR	RC	SVM
1	0.963333333	0.996666667	0.976666667	0.983333333
2	0.9966	0.993333333	0.994466667	0.995533333
3	0.978866667	0.983266667	0.981	0.9878
4	0.872266667	0.8596	0.859666667	0.864933333
5	0.879933333	0.873866667	0.878533333	0.8824
6	0.942666667	0.935333333	0.944666667	0.937333333
7	0.938	0.939333333	0.944	0.936
8	0.867533333	0.862266667	0.868933333	0.861733333
9	0.9906	0.988333333	0.989533333	0.9966
10	0.626933	0.599867	0.549067	0.632533
11	0.597333	0.641333	0.670667	0.664
12	0.677333	0.744	0.721333	0.745333
13	0.605333	0.621333	0.625333	0.593333
14	0.642	0.630667	0.596	0.632667
15	0.644	0.632667	0.594	0.635333
16	0.615333	0.606667	0.574	0.61
17	0.616467	0.595133	0.5444	0.5976

Table A.3: The combined results of 15 participants for the Complete Profiling scenario with the THE syllable

TASK	LDA	LR	RC	SVM
1	0.815(0.076)	0.988(0.016)	0.979(0.022)	0.989(0.019)
2	0.997(0.005)	0.995(0.006)	0.997(0.005)	0.996(0.006)
3	0.987(0.009)	0.987(0.008)	0.988(0.009)	0.988(0.009)
4	0.890(0.058)	0.882(0.051)	0.872(0.05)	0.902(0.051)
5	0.888(0.054)	0.885(0.044)	0.888(0.05)	0.912(0.043)
6	0.959(0.045)	0.958(0.047)	0.961(0.045)	0.959(0.047)
7	0.950(0.042)	0.95(0.043)	0.95(0.041)	0.955(0.034)
8	0.886(0.064)	0.885(0.061)	0.882(0.060)	0.897(0.058)
9	0.999(0.002)	0.999(0.002)	0.999(0.002)	0.998(0.003)
10	0.608(0.093)	0.573(0.087)	0.540(0.074)	0.637(0.088)
11	0.601(0.105)	0.612(0.109)	0.616(0.092)	0.636(0.104)
12	0.72(0.102)	0.716(0.107)	0.725(0.101)	0.716(0.098)
13	0.544(0.093)	0.542(0.089)	0.549(0.084)	0.560(0.096)
14	0.628(0.099)	0.60(0.11)	0.589(0.087)	0.652(0.108)
15	0.627(0.077)	0.588(0.080)	0.575(0.072)	0.637(0.085)
16	0.578(0.081)	0.535(0.082)	0.510(0.08)	0.595(0.089)
17	0.591(0.093)	0.565(0.092)	0.535(0.081)	0.611(0.098)

Table A.4: The combined results of 15 participants for the Complete Profiling scenario with all three syllables

TASK	LDA	LR	RC	SVM
1	NA	NA	NA	NA
2	0.843 (0.169)	0.781 (0.166)	0.853 (0.144)	0.788 (0.159)
3	0.772 (0.182)	0.836 (0.005)	0.783 (0.186)	0.847 (0.024)
4	0.412 (0.108)	0.389 (0.095)	0.432 (0.089)	0.348(0.000)
5	0.567 (0.087)	0.567 (0.087)	0.571 (0.094)	0.567 (0.080)
6	0.662 (0.134)	0.649 (0.108)	0.662 (0.135)	0.674 (0.147)
7	0.625 (0.128)	0.625 (0.114)	0.620 (0.123)	0.622 (0.128)
8	0.465 (0.103)	0.460 (0.095)	0.464 (0.100)	0.492 (0.100)
9	0.866 (0.173)	0.873 (0.167)	0.873 (0.167)	0.813 (0.169)
10	0.132 (0.044)	0.126 (0.037)	0.128 (0.037)	0.067 (0.000)
11	0.254 (0.053)	0.248 (0.057)	0.250 (0.062)	0.248 (0.057)
12	0.240 (0.051)	0.233 (0.043)	0.233 (0.043)	0.236(0.052)
13	0.273 (0.059)	0.264 (0.051)	0.275 (0.059)	0.263 (0.054)
14	0.137 (0.031)	0.135 (0.035)	0.133 (0.031)	0.130 (0.034)
15	0.172 (0.038)	0.168 (0.037)	0.168 (0.040)	0.170 (0.050)
16	0.162 (0.039)	0.161 (0.038)	0.160 (0.040)	0.157 (0.044)
17	0.118 (0.031)	0.110 (0.026)	0.112 (0.027)	0.114 (0.041)

Table A.5: The combined results of 15 participants for the Location Profiling scenario with all three syllables

TASK	LDA	LR	RC	SVM
8	0.795(0.091)	0.326(0.035)	0.323(0.027)	0.33(0.034)
5	0.791(0.095)	0.499(0.06)	0.491(0.042)	0.491(0.037)
6	0.904(0.092)	0.499(0.029)	0.497(0.052)	0.486(0.04)
7	0.908(0.06)	0.508(0.037)	0.499(0.051)	0.494(0.05)

Table A.6: The combined results of 15 participants for the User Profiling scenario with all three syllables

Task No	Task Description
1	Room Classification(Room I1(10 DP) Vs Room O1 DP))
2	Room Classification(Room I1(50 DP) Vs Room O1(15 DP))
3	Room Classification(Room I3(50 DP) Vs Room O1(15 DP))
4	Room Classification(Room I1(50 DP) Vs Room O1(15 DP) Vs Room I2(50 DP) Vs Room I3(50 DP)
5	Room Classification(Room I1(50 DP) Vs Room I2(50 DP))
6	Room Classification(Room I2(50 DP) Vs Room I3(50 DP))
7	Room Classification(Room I1(50 DP) Vs Room I3(50 DP))
8	Room Classification(Room I1(50 DP) Vs Room I2(50 DP) Vs Room I3(50 DP))
9	Room Classification(Room I2 50 DP) Vs Room O1(15 DP))
10	Corner Classification(Room I1(50 DP) Vs Room O1(15 DP) Vs Room I2(50 DP) Vs Room I3(50 DP))
11	Corner Classification (Room I1(50 DP))
12	Corner Classification(Room I2(50 DP))
13	Corner Classification(Room I3(50 DP))
14	Corner Classification (Room I1(50 DP) Vs Room I2(50 DP))
15	Corner Classification (Room I2(50 DP) Vs Room I3(50 DP))
16	Corner Classification (Room I1(50 DP) Vs Room I3(50 DP))
17	Corner Classification(Room I1(50 DP) Vs Room I2(50 DP) Vs Room I3(50 DP))

Table A.7: Legend

B

APPENDIX

B.1. RESULTS FOR DIFFERENT POSITIONS

TASK	LDA	LR	RC	SVM	VC(Hard)	VC(Soft)
2	0.986	0.993	0.993	0.969	0.993	0.993
3	0.953	0.9485	0.948	0.939	0.966	0.9745
4	0.955	0.97	0.955	0.96	0.97	0.97
5	0.983	0.9835	1	0.985	0.9915	1
6	0.7605	0.77	0.7395	0.774	0.774	0.7865
7	0.77	0.75	0.71	0.74	0.74	0.78
8	0.72	0.76	0.67	0.64	0.7	0.77
9	0.79	0.81	0.77	0.92	0.8	0.81

Table B.1: The results of Position 1 (Dataset comprising of 230 datapoints and 3 locations (2 indoor and 1 outdoor)

TASK	LDA	LR	RC	SVM	VC(Hard)	VC(Soft)
2	**1.000**	1.000	1.000	1.000	1.000	1.000
3	**0.983**	0.947	0.974	0.983	0.983	0.983
4	**0.950**	0.950	0.960	0.950	0.970	0.950
5	**1.000**	1.000	1.000	1.000	1.000	1.000
6	**0.973**	0.939	0.948	0.973	0.973	0.973
7	**0.980**	0.940	0.910	0.960	0.980	0.980
8	**1.000**	1.000	1.000	1.000	1.000	1.000
9	**0.960**	0.900	0.920	0.940	0.940	0.960

Table B.2: The results of Position 2 (Dataset comprising of 230 datapoints and 3 locations (2 indoor and 1 outdoor))

B.2. RESULTS FOR DIFFERENT PHONES

TASK	LDA	LR	RC	VC(Hard)	VC(Soft)	VC(Soft)
1	0.989	1	1	1	1	
2	1	1	1	1	1	1.000
3	0.983	0.997333	0.977	0.988667	0.991667	0.983
4	0.99	0.996667	0.99	0.99	0.99	0.950
5	1	0.995333	1	1	1	1.000
6	0.890333	0.884667	0.799667	0.881333	0.913333	0.973
7	0.86	0.876667	0.776667	0.856667	0.883333	0.980
8	0.926667	0.886667	0.893333	0.893333	0.933333	1.000
9	0.873333	0.906667	0.8	0.866667	0.886667	0.960

Table B.3: The results for Different Phones (Dataset comprising of 345 datapoints and 3 locations (2 indoor and 1 outdoor))

Test Data	**Test Data – Speaker C**		**Test Data – Speaker D**		**Test Data – Speaker E**	
Task	Different	Same	Different	Same	Different	Same
2	0.815	0.769	0.954	0.700	0.846	1
5	0.969	1	0.862	1.000	0.769	1
4	0.760	0.5	0.700	0.700	0.900	0.750
3	0.635	0.565	0.643	0.696	0.861	0.730
8	0.400	0.26	0.320	0.360	0.340	0.460
9	0.220	0.36	0.400	0.260	0.460	0.380
7	0.210	0.13	0.400	0.270	0.200	0.310
6	0.815	0.235	0.235	0.365	0.217	0.365

Table B.4: The results for Same and Different Phones (Dataset comprising of 575 datapoints and 3 locations (2 indoor and 1 outdoor))

B.3. RESULTS FOR DIFFERENT AUDIO CONTENT

B.3.1. RESULTS FOR SYLLABLES

TASK	LDA	LR	RC	SVM	VC(Hard)	VC(Soft)
1	0.973333	1	1	1	1	1
2	0.985	1	0.985	1	0.990333	1
3	0.941333	0.923333	0.921	0.924667	0.944	0.95
4	0.95	0.953333	0.94	0.943333	0.946667	0.956667
5	0.979333	0.994333	0.988667	0.995	0.994333	0.994333
6	0.672667	0.667	0.628667	0.611333	0.688	0.714333
7	0.613333	0.68	0.72	0.606667	0.686667	0.753333
8	0.626667	0.64	0.666667	0.633333	0.673333	0.693333
9	0.633333	0.633333	0.63	0.573333	0.663333	0.68

Table B.5: The results for syllable audio content (Dataset comprising of 345 datapoints and 3 locations (2 indoor and 1 outdoor))

B.3.2. RESULTS FOR EXTRACTED SYLLABLES

TASK	LDA	LR	RC	SVM	VC(Hard)	VC(Soft)
1	0.92	0.986667	0.96	0.973333	0.989	1
2	0.989667	0.989667	0.985	0.995	1	1
3	0.915	0.888333	0.903333	0.924667	0.93	0.941667
4	0.9	0.893333	0.903333	0.923333	0.913333	0.93
5	0.95	0.985667	0.981	0.984667	0.981	0.981
6	0.582667	0.536	0.498667	0.577	0.569667	0.621333
7	0.486667	0.526667	0.526667	0.553333	0.573333	0.573333
8	0.52	0.546667	0.54	0.566667	0.553333	0.606667
9	0.53	0.446667	0.463333	0.543333	0.523333	0.56

Table B.6: The results for extracted syllable audio content (Dataset comprising of 345 datapoints and 3 locations (2 indoor and 1 outdoor))

Task No	Task Description
1	Room Classification(Room D(10 DP) Vs Room E15 DP))
2	Room Classification(Room D(50 DP) Vs Room E(15 DP))
3	Room Classification(Room D(50 DP) Vs Room E(15 DP) Vs Room F(50 DP))
4	Room Classification(Room D(50 DP) Vs Room F(50 DP))
5	Room Classification(Room F(50 DP) Vs Room E(15 DP))
6	Corner Classification(Room D(50 DP) Vs Room E(15 DP) Vs Room F(50 DP))
7	Corner Classification (Room D(50 DP) Vs Room F(50 DP))
8	Corner Classification (Room D(50 DP))
9	Corner Classification(Room F(50 DP))

Table B.7: Task description

B.3.3. RESULTS FOR DIFFERENT MESSAGE CONTENT

TASK	LDA	LR	RC	SVM	VC - Hard	VC - Soft
1	0.640 (0.150)	0.880 (0.160)	0.760 (0.196)	0.880 (0.160)	0.883 (0.183)	0.833 (0.211)
2	0.936 (0.109)	0.933 (0.111)	0.933 (0.111)	0.954 (0.092)	0.967 (0.067)	0.950 (0.107)
3	0.986 (0.043)	0.986 (0.043)	0.986 (0.043)	0.969 (0.038)	0.986 (0.043)	0.986 (0.043)
4	0.781 (0.081)	0.738 (0.110)	0.750 (0.111)	0.770 (0.041)	0.775 (0.093)	0.817 (0.092)
5	0.760 (0.111)	0.810 (0.130)	0.790 (0.130)	0.840 (0.080)	0.790 (0.137)	0.840 (0.111)
6	0.860 (0.102)	0.840 (0.136)	0.880 (0.098)	0.870 (0.093)	0.830 (0.078)	0.870 (0.119)
7	0.920 (0.087)	0.900 (0.089)	0.910 (0.094)	0.870 (0.087)	0.890 (0.094)	0.920 (0.087)
8	0.753 (0.112)	0.733 (0.126)	0.747 (0.102)	0.753 (0.027)	0.747 (0.093)	0.780 (0.108)
9	1.000 (0.000)	1.000 (0.000)	0.983 (0.050)	0.985 (0.031)	0.983 (0.050)	1.000 (0.000)
10	0.473 (0.062)	0.485 (0.057)	0.418 (0.075)	0.467 (0.036)	0.469 (0.150)	0.535 (0.100)
11	0.360 (0.233)	0.520 (0.256)	0.520 (0.256)	0.460 (0.102)	0.480 (0.256)	0.560 (0.233)
12	0.640 (0.196)	0.580 (0.227)	0.520 (0.204)	0.600 (0.179)	0.600 (0.179)	0.680 (0.183)
13	0.600 (0.155)	0.560 (0.215)	0.620 (0.108)	0.740 (0.102)	0.660 (0.092)	0.660 (0.092)
14	0.460 (0.162)	0.460 (0.120)	0.410 (0.158)	0.440 (0.107)	0.450 (0.128)	0.500 (0.100)
15	0.560 (0.092)	0.540 (0.102)	0.510 (0.130)	0.560 (0.073)	0.570 (0.135)	0.560 (0.120)
16	0.500 (0.118)	0.440 (0.143)	0.480 (0.166)	0.460 (0.136)	0.480 (0.117)	0.540 (0.128)
17	0.540 (0.128)	0.380 (0.130)	0.393 (0.096)	0.360 (0.106)	0.440 (0.120)	0.493 (0.104)

Table B.8: Results for different audio content (Dataset comprising of 330 datapoints and 4 locations (3 indoor and 1 outdoor))

Task No	Task Description
1	Room Classification(Room I1(10 DP) Vs Room O1 DP))
2	Room Classification(Room I1(50 DP) Vs Room O1(15 DP))
3	Room Classification(Room I3(50 DP) Vs Room O1(15 DP))
4	Room Classification(Room I1(50 DP) Vs Room O1(15 DP) Vs Room I2(50 DP) Vs Room I3(50 DP)
5	Room Classification(Room I1(50 DP) Vs Room I2(50 DP))
6	Room Classification(Room I2(50 DP) Vs Room I3(50 DP))
7	Room Classification(Room I1(50 DP) Vs Room I3(50 DP))
8	Room Classification(Room I1(50 DP) Vs Room I2(50 DP) Vs Room I3(50 DP))
9	Room Classification(Room I2 50 DP) Vs Room O1(15 DP))
10	Corner Classification(Room I1(50 DP) Vs Room O1(15 DP) Vs Room I2(50 DP) Vs Room I3(50 DP))
11	Corner Classification (Room I1(50 DP))
12	Corner Classification(Room I2(50 DP))
13	Corner Classification(Room I3(50 DP))
14	Corner Classification (Room I1(50 DP) Vs Room I2(50 DP))
15	Corner Classification (Room I2(50 DP) Vs Room I3(50 DP))
16	Corner Classification (Room I1(50 DP) Vs Room I3(50 DP))
17	Corner Classification(Room I1(50 DP) Vs Room I2(50 DP) Vs Room I3(50 DP))

Table B.9: Legend

B.3.4. RESULTS FOR SILENT AUDIO CONTENT

TASK	LDA	LR	RC	VC(Hard)	VC(Soft)
Room Classification(Room A(10 DP) Vs Room B(15 DP))	0.960 (0.080)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Room Classification(Room A(50 DP) Vs Room B(15 DP))	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Corner Classification (Room A(50 DP))	0.340 (0.162)	0.380 (0.133)	0.320 (0.117)	0.320 (0.133)	0.400 (0.126)
Corner Classification (Room A(100 DP))	0.500 (0.089)	0.430 (0.040)	0.430 (0.051)	0.510 (0.158)	0.510 (0.158)
Corner Classification (Room A(150 DP))	0.467 (0.084)	0.467 (0.114)	0.487 (0.098)	0.513 (0.166)	0.527 (0.138)
Corner Classification (Room A(200 DP))	0.560 (0.090)	0.530 (0.040)	0.505 (0.043)	0.555 (0.079)	0.590 (0.089)
Corner Classification (Room A(250 DP))	0.581 (0.062)	0.513 (0.038)	0.528 (0.063)	0.547 (0.109)	0.574 (0.089)

Table B.10: Results for silent audio content (Dataset comprising of 265 datapoints and 2 locations (1 indoor and 1 outdoor))

B.4. RESULTS FOR SPEAKER DATA

B.4.1. RESULTS FOR PHONE SPEAKER

TASK	LDA	LR	RC	SVM	VC - Hard	VC - Soft
Room Classification(Room D(50 DP) Vs Room F(50 DP))	0.466667	0.533333	0.483333	0.58	0.483333	0.483333
Corner Classification (Room D(50 DP) Vs Room F(50 DP))	0.266667	0.3	0.166667	0.366667	0.266667	0.266667
Corner Classification (Room D(50 DP))	0.375	0.333333	0.333333	0.5	0.375	0.416667
Corner Classification(Room F(50 DP))	0.148	0.222	0.166667	0.185	0.185	0.166667

Table B.11: Results for phone speaker data (Dataset comprising of 300 datapoints and 2 indoor locations)

B.4.2. RESULTS FOR JBL GO SPEAKER

RECORDING AND TESTING ON SAME DEVICE

TASK	LDA	LR	RC	SVM	VC - Hard	VC - Soft
Room Classification(Room D(50 DP) Vs Room F(50 DP))	0.57	0.68	0.65	0.75	0.6	0.62
Corner Classification (Room D(50 DP) Vs Room F(50 DP))	0.2	0.2	0.26	0.28	0.22	0.22
Corner Classification (Room D(50 DP))	0.3	0.325	0.325	0.425	0.3	0.325
Corner Classification(Room F(50 DP))	0.1222	0.1554	0.1222	0.1666	0.1112	0.1222

Table B.12: Results for JBL GO speaker data (Dataset comprising of 500 datapoints and 2 indoor locations)

TASK	LDA	LR	RC	SVM	VC - Hard	VC - Soft
Room Classification(Room D(50 DP) Vs Room F(50 DP))	0.575	0.75	0.75	0.87	0.725	0.675
Corner Classification (Room D(50 DP) Vs Room F(50 DP))	0.15	0.25	0.2	0.25	0.2	0.2
Corner Classification (Room D(50 DP))	0.25	0.375	0.25	0.4375	0.25	0.3125
Corner Classification(Room F(50 DP))	0.1665	0.25	0.1665	0.2775	0.222	0.1945

Table B.13: Results for JBL GO speaker data - Testing on an unknown speaker/recordee (Dataset comprising of 200 datapoints and 2 indoor locations)

RECORDING AND TESTING ON DIFFERENT DEVICE

TASK	LDA	LR	RC	SVM	VC - Hard	VC - Soft
Room Classification(Room D(50 DP) Vs Room F(50 DP))	0.5875	0.5375	0.5875	0.6675	0.5	0.525
Corner Classification (Room D(50 DP) Vs Room F(50 DP))	0.225	0.325	0.3	0.45	0.275	0.275
Corner Classification (Room D(50 DP))	0.3125	0.34375	0.3125	0.40625	0.25	0.25
Corner Classification(Room F(50 DP))	0.153	0.20825	0.20825	0.25	0.167	0.153

Table B.14: Results for JBL GO speaker data (Dataset comprising of 400 datapoints and 2 indoor locations)

TASK	LDA	LR	RC	SVM	VC - Hard	VC - Soft
Room Classification(Room D(50 DP) Vs Room F(50 DP))	0.525	0.5	0.5	0.65	0.5	0.5
Corner Classification (Room D(50 DP) Vs Room F(50 DP))	0.25	0.25	0.25	0.35	0.25	0.25
Corner Classification (Room D(50 DP))	0.25	0.3125	0.25	0.3125	0.25	0.1875
Corner Classification(Room F(50 DP))	0.139	0.139	0.167	0.167	0.139	0.139

Table B.15: Results for JBL GO speaker data - Testing on an unknown speaker/recordee (Dataset comprising of 200 datapoints and 2 indoor locations)

CONSENT FORMS

INFORMED WRITTEN CONSENT

I Mr. / Mrs. / Ms. _ _____ aged about 25 years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (ONE PLUS 5T) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : Bangalore

Date : 10.02.2021

Name :

Signature:

INFORMED WRITTEN CONSENT

I Mr. /~~Mrs. / M~~s. __ _____ aged about 20 years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (Moto E5) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : Bangalore

Date : 25/02/21

Name :

Signature:

INFORMED WRITTEN CONSENT

I Mr. / ~~Mrs.~~ / ~~Ms.~~ _____ aged about _____ 60 _____ years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (ONE PLUS 6T) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : Bangalore

Date : 10.02.2021

Name :

Signature:

INFORMED WRITTEN CONSENT

I Mr. / Mrs. / Ms. ____ _____ aged about 28 years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (ONE PLUS NORD) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : BANGALORE

Date : 10.02.2021

Name :

Signature:

INFORMED WRITTEN CONSENT

I Mr. / Mrs. / Ms. _ _____ aged about 59 years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (ONE PLUS 3) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : Bangalore

Date : 10.02.2021

Name :

Signature:

INFORMED WRITTEN CONSENT

I Mr. / Mrs (/ Ms.) _____ aged about 29 years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (IOS - I-PHONE) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : Bangalore

Date : 5.

Name :

Signature:

INFORMED WRITTEN CONSENT

I ~~Mr.~~ / ~~Mrs.~~ / Ms. _____ aged about 25 years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (IPHONE 11 PRO.) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : BANGALORE

Date : 08.02.2021.

Name :

Signature:

INFORMED WRITTEN CONSENT

I ~~Mr.~~ / Mrs. / ~~Ms.~~ _ aged about 50 years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (Galaxy A30) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : Bangalore

Date : 27/2/2021

Name :

Signature:

INFORMED WRITTEN CONSENT

I ~~Mr.~~ / Mrs. / ~~Ms.~~ _____ aged about _____ years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (_____) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : BANGALORE

Date : 20/2/21.

Name :

Signature:

INFORMED WRITTEN CONSENT

I Mr. / ~~Mrs.~~ / ~~Ms.~~ _____ aged about 34 years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (SAMSUNG Z FOLD2) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : BANGALORE

Date : 11/FEB/2021

Name :

Signature:

INFORMED WRITTEN CONSENT

I Mr. / Mrs. / Ms. aged about 25 years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (ONE PLUS 6) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : BANGALORE

Date : 11/2/2021

Name :

Signature:

INFORMED WRITTEN CONSENT

I Mr. / Mrs. / Ms. _ _ aged about
25 years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (MOTO G3) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : BANGALORE

Date : 11/02/2021

Name :

Signature:

INFORMED WRITTEN CONSENT

I ~~Mr.~~ / ~~Mrs.~~ / Ms. _____ aged about 21 years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (ONEPLUS 6T) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : BENGALURU

Date : 1 MARCH 2021

Name :

Signature:

INFORMED WRITTEN CONSENT

I Mr. / Mrs. / Ms. _____ aged about _____ years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (SAMSUNG GALAXY A7) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : Bangalore

Date : 1-03-2021

Name :

Signature:

INFORMED WRITTEN CONSENT

I ~~Mr. / Mrs.~~ / Ms. _ aged about 28 years have been explained, in the language best understood by me about the study titled "ENVIRONMENTAL AUDIO CLASSIFICATION".

I have been explained the investigations that will be done during this study. I have no objection to sharing my details and audio recordings recorded on my device (iphone 7) with the investigators of this study. I have been explained that the data may be used for publication/dissertation purposes without revealing my identity.

I understand that my participation in this study is entirely voluntary and I am willing to take part in this study.

Place : BENGALURU

Date : 18-03-2021

Name :

Signature: