Acoustic Sensing for the Detection of Facial Expressions

Project-I (CS47007) report submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Avijit Mandal

(18CS30010)

Under the supervision of

Professor Sandip Chakraborty

Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur

Autumn Semester, 2021-22

November 16, 2021

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: November 16, 2021 (Avijit Mandal)
Place: Kharagpur (18CS30010)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Acoustic Sensing for the Detection of Facial Expressions**" submitted by **Avijit Mandal** (Roll No. 18CS30010) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering is a record of bona fide work carried out by him under my supervision and guidance during Autumn Semester, 2021-22.

Date: November 16, 2021

Place: Kharagpur

Professor Sandip Chakraborty
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur - 721302, India

Abstract

Name of the student: **Avijit Mandal** Roll No: **18CS30010**

Degree for which submitted: **Bachelor of Technology**

Department: **Department of Computer Science and Engineering**

Thesis title: **Acoustic Sensing for the Detection of Facial Expressions**

Thesis supervisor: **Professor Sandip Chakraborty**

Month and year of thesis submission: **November 16, 2021**

Nowadays, we can easily predict the state of facial expression by capturing images and feeding the same through Deep Learning models. When it comes to real-time prediction of facial expression, it might be uncomfortable for a user to keep the camera on while using a particular application for different privacy concerns. Also, cameras with low-resolution can feed images that our model can misclassify.

We propose a method that can predict facial expression without any additional hardware support. We are trying to achieve this by producing Acoustic Signals, which are inaudible to the human ear from commodity devices like smartphones, and recording the reflected signals from the user's face. After that, we can analyze the reflected audio signal to predict the facial expression.

Currently, We prepared a basic android application that can create Chirp Signals and record the same. Also, a baseline Convolution model is designed by training it from JAFFE[7] and CK-plus[10] dataset to predict the facial expression. The following section of this paper will cover problem statements, proposed architecture, current implementation in detail. Finally, we listed out our next plan of action.

Acknowledgements

I would like to thank Dr. Sandip Chakraborty for his precious guidance and support throughout the Project. I would also like to thank Ms. Pragma Kar for her constant support and help at every step of my Project. Furthermore, I would like to thank the Department of Computer Science and Engineering for the facilities and support provided, which were essential for the Project. Many thanks to everyone who made this Project a successful learning experience.

Contents

List of Figures

List of Tables

Abbreviations

OS	**O**perating **S**ystem
RAM	**R**andom **A**ccess **M**emory
CNN	**C**onvolutional **N**eural **N**etwork
PCM	**P**ulse **C**ode **M**odulation
IIR	**I**nfinit **I**mpulsive **R**esponse

Symbols

f	frequency
ϕ	phase of signal

Chapter 1

Introduction

1.1 Acoustic Sensing and Applications

Recent studies have demonstrated the technical feasibility and effectiveness of using acoustic signals for sensing with advancements in wireless and sensing technologies. In the past decades, low-cost audio infrastructures have been widely-deployed and integrated into mobile and Internet of Things (IoT) devices to facilitate a broad array of applications, including human activity recognition, tracking, localization, and security monitoring. The technology underpinning these applications lies in analyzing propagation properties of acoustic signals (e.g., reflection, diffraction, and scattering) when they encounter human bodies. As a result, these applications serve as the foundation to support various daily functionalities such as safety protection, innovative healthcare, and intelligent appliance interaction. Previous works demonstrate how we can use Acoustic Sensing to measure heartbeats[6], finger movements[11] and construct an Indoor Floor Map[1][9].

In this project, we are trying to create an application that leverages the use of Acoustic Sensing to predict the facial expression of an individual.

1.2 Project Idea

The project aims at analyzing smartphone-generated acoustic signals for the purpose of facial expression detection. In this work, we aim at developing a smartphone application, that can produce audio signals of different frequency ranges and record the reflected multi-path signal. Since ultrasound signals with higher frequency range do not interfere with the audible ranges, we use high frequency audio signals of ¿20kHz and capture the unfiltered reflected signal for analysis. By analyzing different features and characteristics of these reflected signals, the application should be able to automatically infer upon the facial expressions of the user, with the help of a deep learning model.

1.3 Challenges

There are multiple challenges in the development of this model:
Firstly, for the correct identification of the changes in facial features, the relevant frequency bin of the reflected signals needs to be identified. Secondly, facial expression change will induce sub-millimeter level displacements which are very intricate, aperiodic and difficult to detect. Thirdly, the features of the reflected signal can be greatly affected by different environmental and embedded noise in form of motions. These needs to be considered and filtered out. Fourthly, the change in expression depends on an individual. Subtle changes are more difficult to detect than the more animated ones. Lastly, it is a matter of research whether all commodity smartphones allow the generation and unfiltered capture of these reflected signals. We tested out our application on different local mobile devices available to us and results were positive till now.

Chapter 2

Problem Statement

The utility of acoustic sensing has been explored in various works, ranging from object tracking to minor activity detection like blinking. This project aims at investigating the nascent yet crucial problem of facial expression classification using acoustic signals. To this extend, camera-based techniques have been used majorly for this purpose, leading to the development of popular facial image databases. However, in this work, we aim to address the problem of facial expression detection by analyzing inaudible audio features. Ultrasound signals have been used for detecting slight movements caused by activities like blinking, proving the capability of audio features to encode sub-millimeter-level body movements. In this work, we aim at detecting the different facial changes, that are likely to be embedded into the reflected ultrasound signals, generated by a commodity smartphone. To the best of our knowledge, this is the first work that aims at solving the challenging problem of facial expression classification through acoustic sensing.

Chapter 3

Problem Motivation

3.1 Problems with Existing Solutions

The main motivation of this work has been driven by the existing challenges of image-based expression detection. Following are a few major problems listed out:

Facial Occlusion Image-based facial expression detection is highly influenced by facial occlusion. Facial occlusion, such as sunglasses, scarves, masks, etc., is one critical factor that affects the performance of face recognition. Unfortunately, faces with occlusion are quite common in the real world, especially in uncooperative scenarios.

Datasets Most of the existing datasets capture enactments of facial expressions, which might vary from natural or slight expression changes in real-life scenarios.

Privacy Concerns It might be uncomfortable for a user to keep the camera on while using a certain application for different privacy concerns. Also the camera resolution may vary device to device, so low resolution cameras can mis-classify.

3.2 Advantages with Acoustic

Ultrasound signals with a higher frequency range (greater than 20KHz) are not affected by the signals with audible frequency ranges. This characteristic encourages the development of such a system, as it will not disrupt the normal audio functionality of the application it will be incorporated in. Most of the recent smartphones support the generation of high-frequency chirp signals that can facilitate the development of this model, thus excluding the requirement of any additional sensor. This will make the system widely acceptable.

Chapter 4

Literature Survey

4.1 Acoustic signals for heartbeat monitoring

Problem Statement Currently, Electrocardiograms (ECGs) and wearable physiological monitors are used to measure vital signs such as heart rate and heartbeat interval. Because these monitoring techniques require direct contact with the patient or are typically uncomfortable, they are too costly and unwieldy to be used regularly.

Solution Proposed In Acousticcardiogram [6] (ACG), the authors proposed the idea of generating inaudible acoustic signals through commonly available devices like smartphones and analyzing the reflected sound to measure the heart rate and heartbeat. The overall goal is to develop a Frequency modulated continuous wave(FMCW) sonar front end that fits commercial audio systems on smart devices, while enabling fine-grained baseband signal processing. Thereafter, they modeled the relationship between FMCW signal phase information and chest movements caused by breathing and heartbeat, and proposed a novel contactless technique that

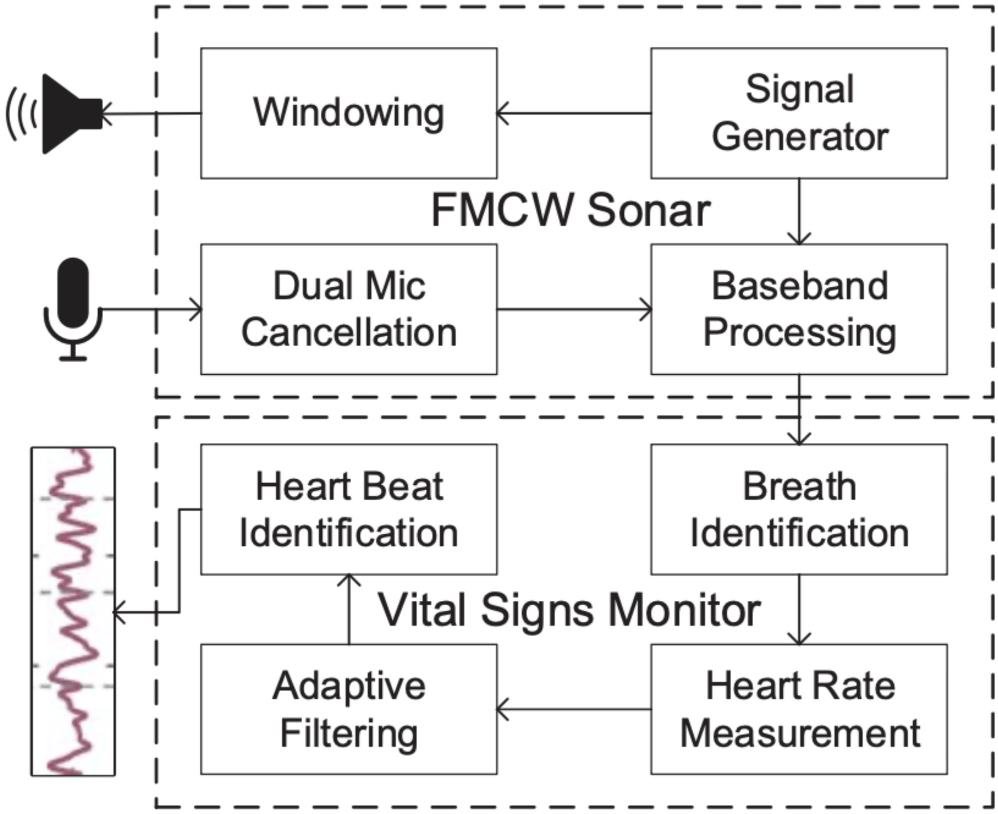


Figure 4.1: Work Flow of ACG

measures vital signs including heart rate and heartbeat interval by monitoring fine-grained FMCW phase information. Figure 4.1 shows the workflow of ACG.

Results Results of ACG's experiments show that a user's heartbeat is accurately monitored, with a median estimation error of 0.6 beats per minute (bpm), and a median estimate error of 19 milliseconds (ms).

4.2 Ultrasound capture from unmodified devices

Problem Statement There is no way for the human ear to hear ultrasonic sounds, as 40kHz is completely inaudible, as well as outside the microphone's range of recording (24KHz). However, it is a challenging and essential task to investigate whether it is possible to record these inaudible sounds without making them audible to the human ear as well and without modifying the recording device or microphone.

Solution Proposed In BackDoor[8], the core idea lies in exploiting non-linearities in microphone hardware. A summary of this idea is that it proposes to design the sound and play it on a speaker such that, after passing through the microphone's nonlinear diaphragm and the power amplifier, the sound creates a "shadow" that is audible. It is possible to modulate shadows in order to carry data bits, allowing acoustic (but inaudible) communication to today's microphones.

Results Experiments demonstrated that BackDoor successfully produced 100 different sounds and 7 individuals confirmed that it was completely inaudible.

4.3 Acoustic signals for floor plan

Problem Statement Developing indoor location-based services (LBS) based on mobile sensing data is challenging due to the lack of digital floor plans. Recent floor plan construction works crowdsource mobile sensing data from smartphone users. These studies typically take a long time (weeks or months) and require a lot of effort, and many rely on images, presenting technical and privacy difficulties.

Solution Proposed BatMapper[1] proposed the idea which explores a previously untapped sensing modality – acoustics – for fast, fine-grained, and low-cost floor plan construction. The authors have developed sound signals that can be detected by heterogeneous microphones on commodity smartphones, and acoustic signal processing techniques to measure the distance to nearby objects..

Results BatMapper is accurate to a distance of 1–2cm in ranges up to around 4m. After a 2–3 minute walk, it generates fine-grained shapes of corridors, detects doors with 92% accuracy, and has proximity errors of 1 to 2m at a 90-percentile

4.4 Acoustic signals and finger tracking

Problem Statement Finger tracking is a task that is useful for a variety of applications including games, human computer interaction, etc. A camera-based technique suffers from occlusion and places a certain amount of constraint on finger location. Commercial tracking devices have the added constraint of limited availability.

Solution Proposed FingerIO [11] is a novel fine-grained finger tracking solution for around-device interaction. FingerIO does not require the finger to be instrumented with sensors and functions even when the finger is occluded. Using Orthogonal Frequency Division Multiplexing (OFDM), smartphones can be converted into an active sonar system that achieves sub-centimeter level accuracy.

Results The evaluation shows that FingerIO can achieve 2-D finger tracking with an average accuracy of 8 mm using the in-built microphones and speaker of a Samsung Galaxy S4.

4.5 Acoustic sensing and blink detection

Problem Statement Many real-life applications use eye blink detection, including Human-Computer Interaction (HCI), drowsy driving prevention, and eye disease detection. Traditional camera-based techniques are promising, but they are hindered in their widespread adoption by several issues, including privacy concerns, strict lighting conditions, and line-of-sight requirements.

Solution Proposed In BlinkListner[5] the authors propose a system to sense the subtle eye blink motion using acoustic signals in a contact-free manner. In this paper, the authors first quantitatively model the relationship between signal variation and the subtle movements caused by eye blink and interference. Then, they propose a novel method that exploits the "harmful" interference to maximize the subtle signal variation induced by eye blinks.

Results on experimental results, BlinkListener provides robust performance with a 95% detection accuracy.

Chapter 5

Proposed Model

In this chapter, the overview and architecture of the system has been described.

5.1 Overview

The application consists of a simple layout that allows the generation of a chirp signal. Currently, the application allows manual selection of the frequency range for testing the optimal range of the generated sound. With the help of the smartphone's speaker, the chirp signals are transmitted periodically. These signals are reflected back to the device from multiple objects, present in the environment. Since the use's face is generally located at a close proximity of the device, the audio signals will also be reflected from the different facial regions of user. Previous studies have shown that these reflected chirps encode even the slightest motion data induced by the body movements of the user. The facial expression change will incorporate significant displacement of facial muscles, as compared to breathing or blinking. As a result, this will be encoded with the features of the reflected chirp. By analyzing different audio features, the application aims at estimating the expression change. The classification of the different facial expressions: Happiness, Neutral, Sad, Contempt, Disgust, Fear,

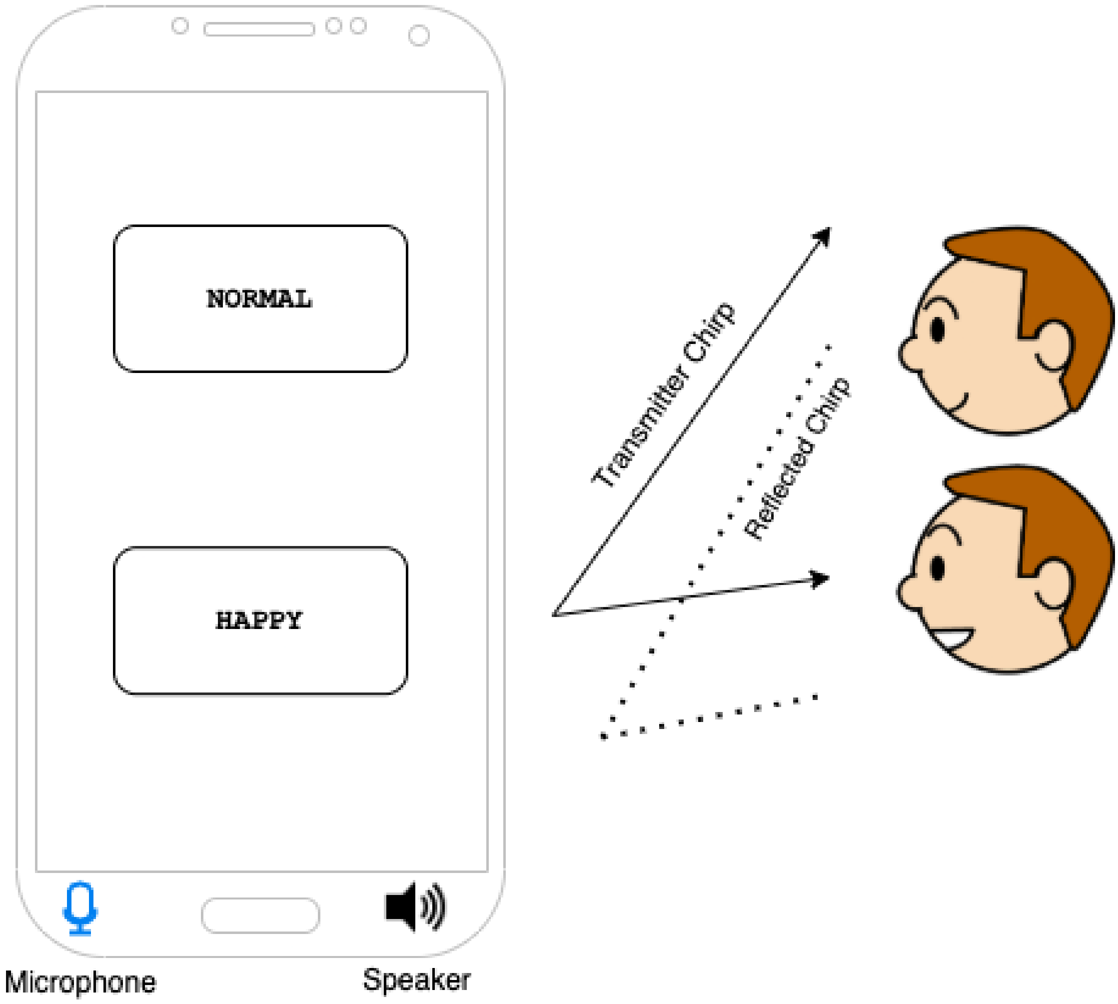


FIGURE 5.1: Overview of the proposed model: Diagram showing how sound signals will be reflected back for analysis and classification of facial expressions

Surprise and Neutral is also facilitated by training a deep Neural Network with the detected audio features. Figure 5.1 shows the overview of the proposed application.

5.2 Architecture

The back-end architecture of the model consists of different sub modules.

5.2.1 Generation of Chirp Signal

The time domain function used to generate linear chirp with frequency ranging from f0 to f1 for a time duration T is given as follows:

$$x(t) = \sin\left[\phi_0 + 2\pi(\frac{ct^2}{2} + f_0 t)\right] \tag{5.1}$$

Where $c = \frac{(f_0+f_1)}{T}$ and ϕ_0 is initial phase of the signal.

The signal was produced for frequency higher than 20000 Hz (ultra sound) and transmitted through the device's speaker. The ambient sound (reflected signals) are also concurrently recorded using the same application.

5.2.2 Feature Extraction

Implemented an IIR high pass filter designed using bilinear transformation into the application. IIR filter design involves the transformation of a continuous-time filter into discrete-time er meeting prescribed specifications. Bilinear transformation is one such transformation technique used to transform continuous-time filters into discrete-time filters. Then converted the filtered sound to a WAV extension file to perform feature extraction from it. To convert PCM to WAV files, we needed to add the relevant headers to the PCM audio. We used JLibrosa, a java library for audio processing, to extract the frequency domain features of the recorded audio. We compute the following features:

1. Mel-frequency cepstral coefficients (MFCC values): MFCC is a type of non-linear spectrum of the spectrum. Formally, MFC represents the short-term power spectrum, based on the cosine transform of the log power spectrum of audio.

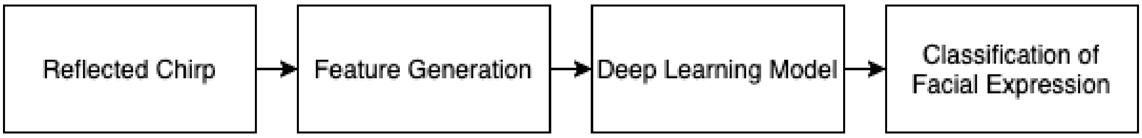


FIGURE 5.2: The proposed architecture: processing pipeline

2. Mel Spectrogram values: It is a spectrogram that is represented in Mel-scale. Mel-scale is a scale of pitches perceived by humans as equally spaced. A frequency of f Hz can be converted into a Mel scale using the following.

$$m = 2599 \log_{1\,0}\left(1 + \frac{f}{700}\right) \tag{5.2}$$

3. Short-time Fourier transforms (STFT values): STFT is used to determine the sinusoidal frequency and phase content of a section of the audio signal.

4. Inverse Short-time Fourier transforms (ISTFT values).

5.2.3 Deep Learning Model and Classification of facial expressions

In these submodules, a deep learning model is used to learn the different audio features and the encoded pattern pertaining to the differences in the facial expressions. Typically a Convolutional Neural Network will be employed for this purpose.

Figure 5.2 depicts the overall processing pipeline of the proposed model.

Chapter 6

Implementation

In this chapter, the current state of implementation has been discussed.

6.1 Device

We tested our application Realme X, which has the following specifications:

Table 6.1: Device Specifications

Row	Specification	Value
1	OS	Android 9 Pie
2	RAM	4 GB
3	Processor	Qualcomm Snapdragon 710

6.2 Implementation Details

This sub-section discusses the implementation of the proposed model and an image based expression classifier, as a baseline approach for comparison.

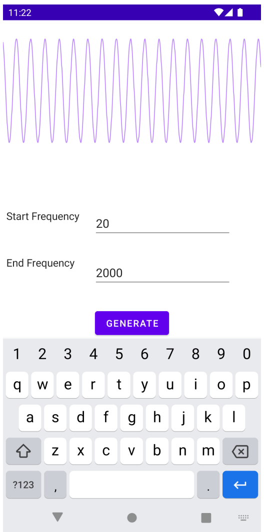


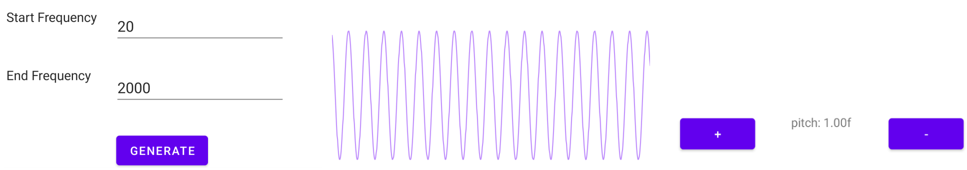
FIGURE 6.1: Overview of the Application

6.2.1 Implementation of the Proposed Model

The major work done in implementing the project was focused on generating an android application, capable of producing ultrasound linear chirp signals and recording the reflected signal along with ambient noise to extract various features from it. Figure 6.1 shows the interface of the developed android application.

Chirp Generation The sampling rate of the chirp signal is taken as 44100 Hz and the duration is 5 sec. The signal has been encoded using PCM and `AudioTrack`[4] API was used to play the signal on an android device. The `AudioTrack` class manages and plays a single audio resource for Java applications. It allows streaming of PCM audio buffers to the audio sink for playback. This is achieved by "pushing" the data to the `AudioTrack` object using one of the `write(byte[], int, int)`, `write(short[], int, int)`, and `write(float[], int, int, int)` methods. PCM is a method used to digitally represent sampled analog signals. It the standard format to store audio signals in computer. Audio track class is used to play audio resources in Java applications. It allows streaming of PCM audio files for playback in applications.

Pitch Controller and Visualization The function for increasing/decreasing the pitch of the signal has been added, where pitch changes by 0.05 on every click. As the ultrasound frequency cannot be heard, a time-domain visualization of the signal has been incorporated into the application to verify the working of the signal generator. Figure 6.2 shows how, the developed application visualizes the signal as a Time domain function and the pitch controller.



(a) Generate Parameters (b) Time-domain visualization (c) Pitch Controller

FIGURE 6.2: Visualization of generated Chirp and Pitch Controller

Recording The application also records the reflected signal along with surrounding noise which is then filtered to obtain high-frequency components. The Recording `AudioRecord` [2] class which is provided by `android.media` library has been used for this purpose. The source of recording is set to device's microphone.

Audio Feature Extraction The filtered sound was converted to .WAV extension file to perform feature extraction from it. To convert PCM to WAV file we needed to add the relevant headers to the PCM audio. We used `JLibrosa`, which is a java library for audio processing, to extract the frequency domain features of the recorded audio.

Storage of Audio Features After extracting the data, a feature for storing the features in form of a string in the `SQLite` database is implemented. For saving the data, a `DatabaseHelper` is implemented which extended `SQLiteOpenHelper` [3].

`OnCreate` method implemented which creates a table named `features` with columns, as shown in Table 6.2. In this table all the values are stored as string and arrays are

TABLE 6.2: Database Descriptions

name	type	notnull	pk
ID	String	1	1
noOfFrames	Integer	0	0
sampleRate	String	0	0
noOfChannels	String	0	0
STFTComplexValues	String	0	0
invSTFTValues	String	0	0
MFCCValues	String	0	0
MeanMFCCValues	String	0	0

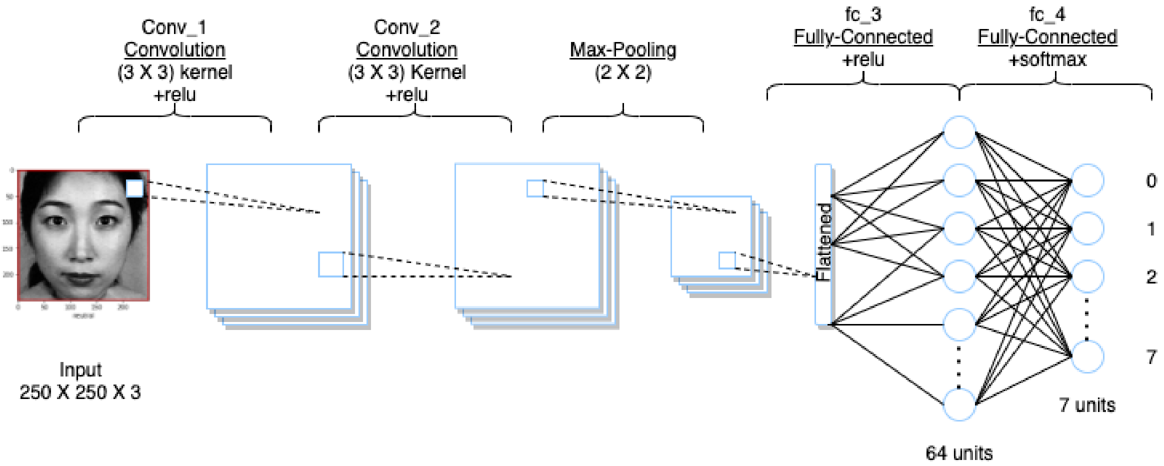


FIGURE 6.3: Basic Block Diagram of the CNN model

stored in comma(,) separated strings.

`addData` method takes all these parameters and store the values in the features table.

6.2.2 Implementation of the Baseline Model

We trained a Convolutional Neural Network (CNN) model (shown in Figure 6.3) to detect the facial expression from a person's image. The model has two convolution layers, each having 64 filters and a kernel size (3, 3). The second convolution layer is followed by a max-pooling, which has a pool size of (2, 2) The steps and dataset descriptions are as follows:

1. We created a combined dataset using JAFFE[7] (Japanese Female Facial Expression) and CK-plus[10] (extended Cohen Kanade). JAFFE data set consists of 213 images, and CK-plus consists of 981 images.

2. Detected the face for each of the images using `open-cv`'s pre-trained `haarcascade_frontalface_default` classifier.

3. Cropped the images around their face boundary and resized every image to same size if 250X250.

4. Trained the CNN model using 80% of the processed dataset and tested it using 20% remaining dataset.

Results We achieved a training accuracy of 97.35% and a testing accuracy of 78.85%.

Chapter 7

Future Work

As a part of the future work, the following tasks will be performed:

1. The phasor diagram, based on the reflected Chirps will be analyzed, to investigate whether it can encode the expression change instances.
2. A pilot study will be conducted to analyze the nature of data in real world.
3. The appropriate Deep Learning model will be developed for learning the audio features for prediction.
4. A thorough Study will be conducted for testing the accuracy of the model.

Chapter 8

Conclusion

The proposed model aims at addressing the challenging problem of facial expression detection and classification through acoustic sensing. The model uses ultrasound signals transmitted and captured by commodity smartphones, without the use of any external sensors. The model is currently under development and is showing promising results in terms of feasibility. In future, we aim at analyzing the full capability of the model through experimental analysis. The aim of the model is to address the existing challenges in image based expression classifications and benefit a wide range of users. The functionality of this application can be extended to attention estimation of learners during online classes, by assessing the rate of expression change, HCI for physically disabled users, Game based learning for learners with cognitive challenges and so on.

Bibliography

[1] Ruipeng Gao Fan Ye Bing Zhou, Mohammed Elbadry. Batmapper: Acoustic sensing based indoor floor plan construction using smartphones. In *MobiSys '17: Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017.

[2] Android Developer Documentation. Audiorecord, android developers. 2018.

[3] Android Developer Documentation. Audiotrack, android developers. 2018.

[4] Android Developer Documentation. Sqliteopenhelper, android developers. 2018.

[5] Lei Wang Jie Xiong Jialin Liu, Dong Li. Blinklistener: "listen" to your eye blink using your smartphone. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2021.

[6] Fu Xiao Yue Zheng Yi Zhang Zheng Yang Yunhao Liu Kun Qian, Chenshu Wu. Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018.

[7] Miyuki; Gyoba Jiro Lyons, Michael; Kamachi. The japanese female facial expression (jaffe) dataset. 2018.

[8] Romit Roy Choudhury Nirupam Roy, Haitham Hassanieh. Backdoor: Making microphones hear inaudible sounds. In *MobiSys '17: Proceedings of the*

15th Annual International Conference on Mobile Systems, Applications, and Services, 2017.

[9] Armin B. Cremers Pascal Bihler, Paul Imhoff. Smartguide – a smartphone museum guide with ultrasound control. In *The 8th International Conference on Mobile Web Information Systems (MobiWIS)*, 2011.

[10] Takeo Kanade Jason Saragih Zara Ambadar Patrick Lucey, Jeffrey F. Cohn. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Disney Research, 4615 Forbes Ave, Pittsburgh, PA 15213*, 2010.

[11] Desney Tan Shyamnath Gollakota Rajalakshmi Nandakumar, Vikram Iyer. Fingerio: Using active sonar for fine-grained finger tracking. In *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.