
Software Engineering Department

Braude College

# Capstone Project Phase B

# SoundSOS

23-1-D-20

Supervisors: Dr Dan Lemberg, Mrs. Elena Kramer

Daniel Ohayon: Ohayon580@gmail.com
Tal Levinzon: Tallevinzon@gmail.com

GitHub URL - https://github.com/DanielOhayo/SoundSOS

# Contents

# Abstract

The growing number of cases of verbal and physical violence together with the increasing awareness by society creates an important need of providing solutions to these situations.

Currently, the treatment for this is not immediate, and in most of the cases there is missing evidence. This makes violent people feel comfortable with this situation and continue to do these terrible things. We present an application that solves the issue that mentioned above. First, we presented the CNN algorithm and its analysis and then presented our solution with the finished product. The application includes the features of unique voice recognition, emotion voice recognition and provide a solution for distress situation. Also, the application will be available for now in android and later will be available on iOS.



Keywords: voice, recognition, distress, recording, CNN, automatic.


# 1. Introduction

The first monograph on the expression of emotions in animals and humans was written by Charles Darwin in the last century. After that psychologists have gradually accumulated knowledge in this field. With the development of the technology, AI researchers made contributions in the following areas: emotional speech synthesis (Canh, 1989; Murray and Arnott, 1993), recognition of emotions (Dellaert et al., 1996), and using agents for decoding and expressing emotions (Tosa and Nakatsu, 1996). [2]

In the context of the emotions mentioned above, cases of distress have become more common, leading to poor quality of life and, in severe cases, death. A case for example is dangerous violence between couples.
Addressing emergency situations typically involves calling the police or using an emergency button. However, this approach can be problematic as it may not always be possible for the person in distress to access the device in a timely manner. Time is a crucial factor in responding to emergency situations, and alternative solutions are needed to ensure a swift response.

Voice recognition technology has the potential to revolutionize the way we recognize and respond to distress. By analyzing the pitch, tone and rhythm of a person's voice, voice recognition systems can accurately detect when someone is in distress, even if they are unable to communicate it verbally. The development of voice recognition to detect

distress brings with it many potential applications to improve people's lives and help in crisis situations.

The aim of our application is to give a sense of security in any place. We will focus on the domain of voice recognition, which is a promising way to detect as soon as possible distress. In the context of our application, the primary tasks of the voice recognition are the followings:

- to monitor the users in a moment of distress in real time via application.
- to inform the relevant parties as soon as possible (the application will call automatically in the emergency number with recording message that contains the user details).
- to provide a documentation of the distress situation.

However, there are still important challenges to overcome with our implementing.
The conditions that challenging because of ambient noise, reverberation, distortion, and the acoustical environment influence.

# 2. Related Work and Background

A similar study conducted earlier looks at the recognition of emotions based on vocal cues, specifically each of our emotions (joy, sadness, anger, and fear). The studies have shown that the voice is indeed a powerful source of information about emotions and that it is difficult to "disguise" them. Recognizing emotions from vocal cues involves identifying the emotions that a person expresses through his or her voice. This can be done through a variety of techniques, such as analyzing the pitch, volume, and rhythm of a person's voice, as well as other acoustic features. It can be useful in a variety of applications, such as customer service, virtual assistants, and mental health assessment. Research in this area is ongoing, and there are still some challenges to overcome, such as developing methods that are suitable for different languages and accents. [4] We focus on the case that can be solved with this tool. Another work is a CNN proposed by Varun et al. that predicts the mood of the animal based on the sounds. Their model was 80% accurate in predicting the mood in real time. [11]

Also, this section is dedicated to the description of the history, causes and current situation.

## *2. 1 Verbal violence*

Verbal violence has numerous effects on organizations and on the health of victims. Several studies have emphasized the need to consider victim characteristics, particularly gender, to better understand the prevalence of verbal violence in the workplace. In fact, study results

are contradictory, with some showing that women are more at risk, while others indicate that men are more at risk. These differences could be explained in part by other factors that affect the prevalence of workplace violence, such as occupational domains and job characteristics. Therefore, the purpose of this literature review was to describe the prevalence of verbal violence by gender in different occupational settings. [5]

### *2.1.1 Arguments in couples*

Verbal content of arguments in couples with a violent husband-Based on self-reports of violent arguments, there were no wife behaviors that successfully suppressed the husband's violence once it began; furthermore, the husband's violence escalated in response to both nonviolent and violent wife behaviors, whereas the wife's violence escalated only in response to husband violence or emotional abuse. Only wives were fearful during violent and nonviolent arguments. Observation of nonviolent arguments in the laboratory revealed that both violent husbands and their wives (DV) were angrier than their maritally disturbed but nonviolent (DNV) counterparts. As predicted, only DV men differed from their DNV counterparts on the more provocative anger codes. However, DV wives were as verbally aggressive toward their husbands as DV husbands were toward their wives. [6]

### *2.1.2 Reasons to call the police*

Fear. Usually, the person who was the target of the verbal violence called the police. Fear is identified as a correlate of risk. Fear was heightened when it was accompanied by attempts to control, property damage, threats, and clear signs of escalation. The occurrence of fear in situations indicates that even when no criminal event appears to have occurred, it still needs to be taken seriously. [7]

## *2.2 Voice recognition*

Voice recognition technology (VRT) has been touted many times as the next "killer app." However, many have tried this technology and put it aside, promising to return to it when the technology improves. As a result, it is easy to understand why this technology has not become more widely adopted. While VRT has been widely accepted by people with disabilities that prohibit or make it difficult for them to type, the vast majority of people who can type continue to use their keyboard skills. Keyboard skills developed over years or decades were considered sufficient or superior to VRT. Because VRT applications require hardware that other software does not, such as a good sound card, a microphone, additional

software, and a certain amount of training, most users will not try them without very compelling reasons [8].

### *2.3 Machine learning*

The study and computer modeling of learning processes in their many manifestations is the subject of machine learning. Currently, the field of machine learning is divided into three main research areas: (1) Task-oriented studies - the design and analysis of learning systems to improve performance on a specific set of tasks, also known as a technical approach. (2) Cognitive simulation - the study and computer simulation of human learning processes. (3) Theoretical analysis - the theoretical exploration of the space of possible learning methods and algorithms, regardless of the application domain. An equally fundamental scientific goal of machine learning is the exploration of alternative learning mechanisms, including the discovery of different induction algorithms, the scope and limitations of particular methods, the information that must be available to the learner, the problem of dealing with imperfect training data, and the development of general techniques that are applicable in many task domains. [9]

#### *2.3.1 Machine learning groups*

Machine learning algorithms are divided into the following groups:

Supervised learning - the various algorithms produce a function that maps inputs to desired outputs. A standard formulation of the supervised learning task is the classification problem: the learner must learn (approximate the behavior of) a function that maps a vector to one of several classes by considering several input-output examples of the function.

Unsupervised learning - models a set of inputs: labeled examples are not available.
Semi-supervised learning - combines both labeled and unlabeled examples to create an appropriate function or classifier.
Reinforcement learning - the algorithm learns a strategy for how to act given an observation of the world. Each action has some effect on the environment, and the environment provides feedback that guides the learning algorithm.
Transduction - similar to supervised learning, but does not explicitly construct a function, instead trying to predict new outputs based on training inputs, training outputs, and new inputs.

Learning to learn - where the algorithm learns its own inductive alignment based on previous experience. [10]

### *2.3.2 The learning process*

The learning process in a simple machine learning model is divided into two steps:
Training and testing.
Training process: samples in the training data are taken as input in which features are learned by the learning algorithm or learner and build the learning model.
Training is the most important step in machine learning. You pass the prepared data to your machine learning model to find patterns and make predictions. The result is that the model learns from the data so that it can accomplish the task at hand. Over time, the model gets better and better at making predictions through training.
Test Process: The learning model uses the execution engine to make the prediction for the test or production data.
After you train your model, you need to verify how it performs. This is done by testing the performance of the model against previously unseen data. The unseen data is the test set into which you previously split our data. If the tests are run on the same data that was used for training, you will not get an accurate measurement because the model will already have become accustomed to the data and will find the same patterns in it as before. The result is disproportionately high accuracy.
When you use test data, you get an accurate measure of your model's performance and its speed.

## *2.4 CNN*

The convolutional Neural Network is a technology in advanced deep learning that can achieve high accuracy in image recognition. Applying a convolution operation is a common technique used in computer vision. CNN contains several layers where each layer performs a specific transformation function. The first layer is a convolutional layer that extracts 10 features from the input. In the next stage, by learning the image features from the input, the convolutional layer will be able to keep the relationship between the pixels. The next layer is the pooling layer. The functions of the pooling layer reduce the number of parameters when the image is too large. There are few pooling functions, for example, average pooling, max pooling, and sum pooling. Each one of them reduces the dimensions of each map but preserves important information. The above steps can repeat several times, depending on the model. All the steps we explained so far assemble the feature learning stage. The last step is to create a fully connected layer which means we flatten the matrix (output from the last layer) into a single vector. The overall architecture of a CNN typically consists of several types of layers:

Input layer: This layer receives the raw input data and passes it through to the next layer.

Convolutional layer: This layer applies a convolution operation to the input data, which involves sliding a small window (called a filter or the kernel) over the input and performing element-wise multiplications and summations to extract features from the data. The convolutional layer also introduces nonlinearity into the network using an activation function.

Pooling layer: This layer reduces the size of the input by applying a down-sampling operation, such as max pooling or average pooling. This helps to reduce the computational cost of the network and can also help to reduce overfitting.

Fully connected (dense) layer: This layer combines the features extracted by the previous layers and uses them to make predictions. The fully connected layer typically has many neurons and is connected to every neuron in the previous layer.[14]

### *2.4.1 Voice command prediction CNN model*

CNN has also used in the human voice classification where the human voice signals are taken as the input and converted into images to feed to the CNN and the output will be the classification of the human voice. [11]

## *2.5 DenseMap*

DenseMap is a machine learning algorithm that can be used for a variety of tasks, including images and audio recognition. It is not specifically designed for audio recognition, but it is possible that you could use DenseMap for this purpose by converting the audio data info a format that DenseMap can process, such us a spectrogram or Mel-Frequency Cepstral Coefficient (MFCC) representation. These types of representation convert audio data into a visual format that can be processed by image recognition algorithms like DenseMap.

## *2.6 LSTM - Long-Short-Term-Memory*

Long short-term memory architecture is the state-of-art model for sequence analysis since it uses memory cells to store information that can exploit long range dependencies in it.[15]

Architecture 3 layers:

1. input
2. forget
3. output

Audio recognize Model:

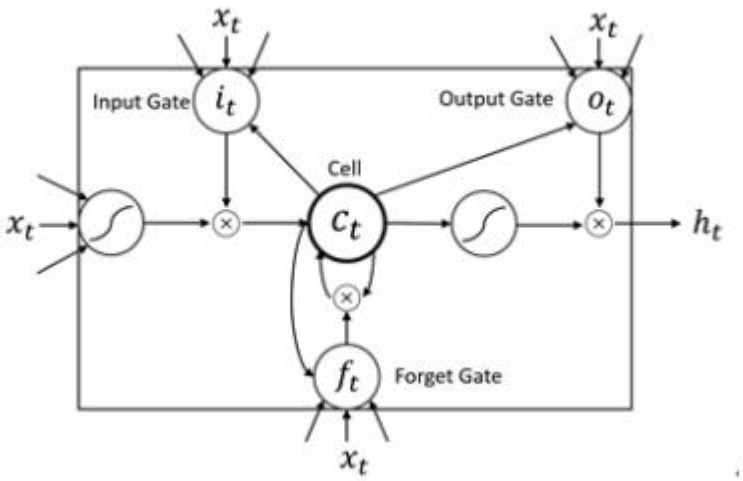



Figure 1.
Audio recognize Model

*2.7 RNN- Recurrent neural network*

Recurrent neural network (RNN) is a type of neural network that is designed to process sequential data, such as time series or natural language. They are able to capture the dependencies between different elements in the sequence by using feedback connections that allow the output of the network at one time step to influence the input at subsequent time steps. RNN can process an entire sequence of audio samples at once, rather than processing one sample at a time like a traditional feedforward neural network. RNNs are able to do this by introducing "memory" into the network in the form of hidden states, which pass from one time step to the next. This allows the network to use information from previous time steps when processing the current time step. There are many different types of RNN architectures that have been developed for speech recognition tasks, such as long short-term memory (LSTM) networks. These architectures are able to effectively capture long-range dependencies in the input data. [15]

## 2.8 Spectrogram

A spectrogram is a visual representation of the frequency spectrum of a signal as it varies over time. It is often used to analyze audio signals, for example to analyze the melodies of a song or to identify the presence of specific words or phrases in a recording. To create a spectrogram, the signal is divided into overlapping time frames, and the frequency content of each frame is represented by a set of vertical bars or "bins" on a graph. The height of each bar indicates the intensity of the frequency component at that particular time, and the color of the bar may be used to indicate the phase or any some other aspect of the signal. Spectrograms can be useful for identifying patterns and structures in signals that are not easily apparent in the raw data. [16]

## 2.9 MFCC - Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCCs) are a feature used in speech and speaker recognition. MFCCs represent the power spectrum of a single frame of audio, but it is believed that speech also contains information on the temporal evolution of the MFCC coefficients. To extract MFCCs from an audio signal, the input is first divided into short frames. The periodogram estimate is then calculated for each frame, and a Mel filter bank is applied to the power spectra. The logarithm of the filterbank energies is taken, and the resulting values are transformed using the discrete cosine transform (DCT). The DCT coefficients 2-13 are retained, while the rest are discarded. This process is shown in Figure 3. MFCCs can improve speech recognition performance by a significant amount when combined with the primary feature vector. [17]

# 3. The Process

### *3.1 Recognition voices*

The audio will be translated to a matrix which will represent an image. The system activates CNN algorithm to train itself to recognize:
*Unique voice* - the user voice.
*Emotions voice* - classification according to emotion such as happiness, sadness, distress etc.
Distress voice - the system will respond to distress voice.

Choosing algorithm - Initially we did research about deep learning algorithms.
First, our focus was on the Neural Network (NN) that is commonly used when creating a supervised machine learning model which is inspired by neuron connections in the human brain.
We have chosen Convolutional Neural Network (CNN) as our algorithm because we are familiar with this algorithm already and we are aware of his abilities and of his efficiency. Also, we did deep research about CNN, and we were impressed by his productivity.

### *3.2 Database selection*

In the past, relational databases were used in a large scope of applications due to their rich set of features, query capabilities and transaction management. Relational databases provide good support for structured data management [12]. However, recent development in IT brings forward big data, featuring extremely large volume and variety in data type and structures, and relational databases are difficult to handle due to the strict constraints on data structure and data relations, and so on [13]. They are not able to store and process big data effectively and are not very efficient to make transactions and join operations [12].

Recently, a new paradigm emerged, NoSQL databases, to overcome some of these problems, which are more suitable for the usage in web environments [13] . NoSQL databases, including HBase, MongoDB, Cassandra, etc., are receiving popularity for their capability in dealing with large amounts of complex data in various structures [12].

### *3.3 Dataset selection*

Unique voice recognition:

TIMIT: Is a corpus of American English speech recordings for use in acoustic-phonetic studies and speech recognition system development. It includes 630 speakers from 8 dialects, each reading 10 sentences. The corpus includes transcriptions and a speech waveform file for each utterance, as well as test and training subsets and searchable information. It was created by

MIT, SRI, and Texas Instruments, with speech recorded by TI, transcribed by MIT, and prepared for CD-ROM production by NIST. All transcriptions have been hand-verified. [17]

*Recognition of emotions:*

RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a dataset of audio and video recordings of actors expressing various emotions. The dataset contains actors speaking and singing in different emotional states, such as happy, sad, angry, and neutral. The recordings are intended for research in affective computing and can be used to train and test machine learning models for speech and emotion recognition. [18]

Berlin: The Berlin dataset, which contains expert annotated speech data from four different speakers, was used to evaluate the performance of the proposed speech emotion recognition system (SER). The dataset contains audio files labeled with one of seven emotions: Neutral, Fear, Anger, Joy, Sadness, Disgust, and Boredom. To obtain results, the dataset was divided into a training set (75%) and a test set, and five-fold cross-validation was performed. [16]

SAVEE: This dataset contains around 500 audio files recorded by 4 different male actors. It contains 4 emotions - anger, happy, sad, and natural. [18]

We choose to use two datasets, the first is RAVDESS because the English language and the large size of audio files can lead to better results.
The second one, SAVEE, because it contains 4 emotions that will help us sharpen the difference between these emotions and the emotion of fear.

## Process challenge

The challenges we faced during the project are:

- Identify unique voices.
- Classification of the voices according to emotions.
- Usage unknown database.
- Make it available by phone.
- Creating a user-friendly application.
- Creating an automation record.

# Methodology and Development Process

For development we chose to go with the Agile methodology which we find to be very fitting to our use case.
This methodology is a project management approach that emphasizes flexibility and rapid response to change.

In this method we work according to iterations:

1. Using a CNN algorithm as a method to identify unique voices.
2. By using a unique voice, activate CNN algorithm as a method to create voice classification according to emotion.
3. Make an application to record when you recognize a distress voice.
4. Get permission from the user phone to access location service.
5. Make an application call automatically to the emergency number.
6. Do some testing.
7. Applying change to any problem that arises through the result.
8. Creating a user's screen with a friendly GUI.
9. Applying the application to be accessible on a phone.

According to the agile method, we work in short, iterative cycles called “sprint". The sprint will be one-two weeks and we will meet every day in a meeting called “daily”.

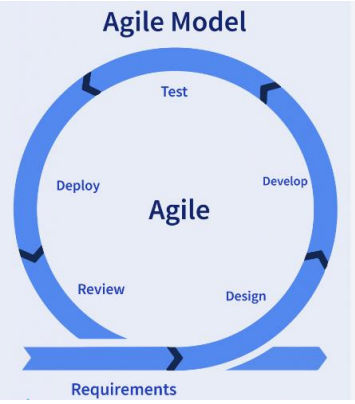



Figure 2.
The agile model

# 4. Product

*4.1 Requirements:*

## Functional:

1. The system recognizes the user's voice.
2. The system knows only the user's voice.
3. The system will know classified user’s voices by emotion.
4. The system activates when it recognizes distress.
5. The system can be turned off/on.
6. The system enables recording audio.
7. The system allows users to define emergency number.
8. The system allows the user to change part of the setting.
9. The system can access location information.
10. The system enables auto calls.

## Non-functional:

1. User voice in the system will be learned by the user.
2. The turn off/on will be by button click.
3. The audio will be saved on the system's DB.
4. The setting contains location, username, emergency number.
5. Pre-recorded message will be sent to the emergency number.
6. The calls will be when a distress situation occurred.
7. The calls will be to the emergency number.

*4.2 Architecture overview:*

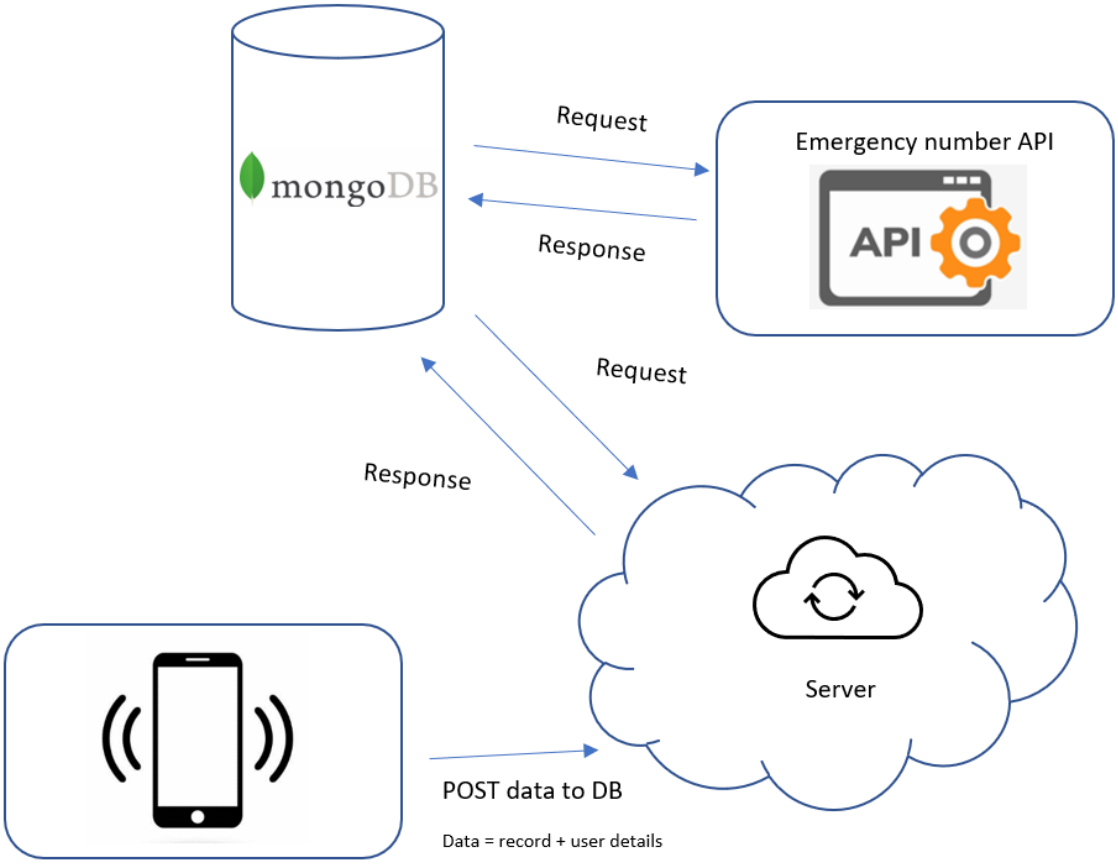



Figure 3.
Software architecture Diagram

In the diagram above, there is the architecture of the project. We will use visual code to write the python code for the deep learning, write java script code for the server and by using the mongoose library, we will connect the database which will be built in the MongoDB.

*4.2 .1 CNN Architecture for biometric recognition*

The proposed CNN model consists of 29 layers: 8 convolutional layers (2 of them are fully connected), 7 Activation function, 5 padding layers, 6 batch normalization layers, 1 pooling layer, 1 reshape layer and 1 Lambda layer.
The Activation function is followed by rectified linear units (ReLU) and the last convolutional layer is a fully connected layer. The convolutional layers include 2-dimensional, and the pooling layer is global average pooling layer that averages the spatial dimensions of the previous outputs. Also, we have the normalization layer that normalizes the previous output and eventually we have the lambda layer which represents a custom operation applied to the outputs.

*4.2 .2 CNN Architecture for emotion recognition*

The proposed CNN model consists of 18 layers: 6 convolutional layers, 7 Activation function, 2 Dropout, 1 Max pooling, 1 Flatten and 1 Fully connected (dense) layer. The Activation function is followed by rectified linear units (ReLU) and the last convolutional layer is a SoftMax layer. The input of the network is a 216 x 1, one-dimensional convolutional layer. Layer C1 has 256 kernels of size 5 x 5 which are applied in a stride setting of 1 pixel. It is followed by rectified linear units (ReLU). ReLU acts as activation functions instead of the typical sigmoid functions which improves efficiency of the training process. Layer C2 has 128 kernels of size 5 x 5, and they are applied to the input with stride 1. It is followed by rectified linear units (ReLU) and a Dropout layer with 0.1 dropout rate and a max pooling layer of size 8 x 8. The similar flow until the latest layers, flatten layer which convert the multidimensional input data into a one-dimensional vector and Dense layer with 10 neurons. It is followed by Softmax.

| Layer (type) | Output Shape | Param # |
|---|---|---|
| conv1d_1 (Conv1D) | (None, 216, 128) | 768 |
| activation_1 (Activation) | (None, 216, 128) | 0 |
| conv1d_2 (Conv1D) | (None, 216, 128) | 82048 |
| activation_2 (Activation) | (None, 216, 128) | 0 |
| dropout_1 (Dropout) | (None, 216, 128) | 0 |
| max_pooling1d_1 (MaxPooling1 | (None, 27, 128) | 0 |
| conv1d_3 (Conv1D) | (None, 27, 128) | 82048 |
| activation_3 (Activation) | (None, 27, 128) | 0 |
| conv1d_4 (Conv1D) | (None, 27, 128) | 82048 |
| activation_4 (Activation) | (None, 27, 128) | 0 |
| conv1d_5 (Conv1D) | (None, 27, 128) | 82048 |
| activation_5 (Activation) | (None, 27, 128) | 0 |
| dropout_2 (Dropout) | (None, 27, 128) | 0 |
| conv1d_6 (Conv1D) | (None, 27, 128) | 82048 |
| activation_6 (Activation) | (None, 27, 128) | 0 |
| flatten_1 (Flatten) | (None, 3456) | 0 |
| dense_1 (Dense) | (None, 10) | 34570 |
| activation_7 (Activation) | (None, 10) | 0 |

Figure 4.

Architecture of emotions voice recognize.

*4.3 Project Design*

Module 1:  Pre - processing

1. Learning level: To be able to recognize the distress in the voice after an incident that caused it, we need to be able to recognize specific sound waves.
   Input = audio, creating a DB according to the neural network training. The DB will contain labeled examples that we can use as reference.
2. Implementation level: input = DB from the above and integrate into the system.

Module 2: Main Flow (take place in the cloud)

The model should be able to detect distress from the voice analysis.

- Input = voice audio, apply the CNN algorithm that analyzes the situation and makes a prediction (True/False).

Module 3: Post-Process

- Waiting for True if true turn on the solution of the system.

Module 3.1: Audio Recording

- Each recording is saved as a file and sent to the server.

- **Diagrams:**

Use-Case Diagram

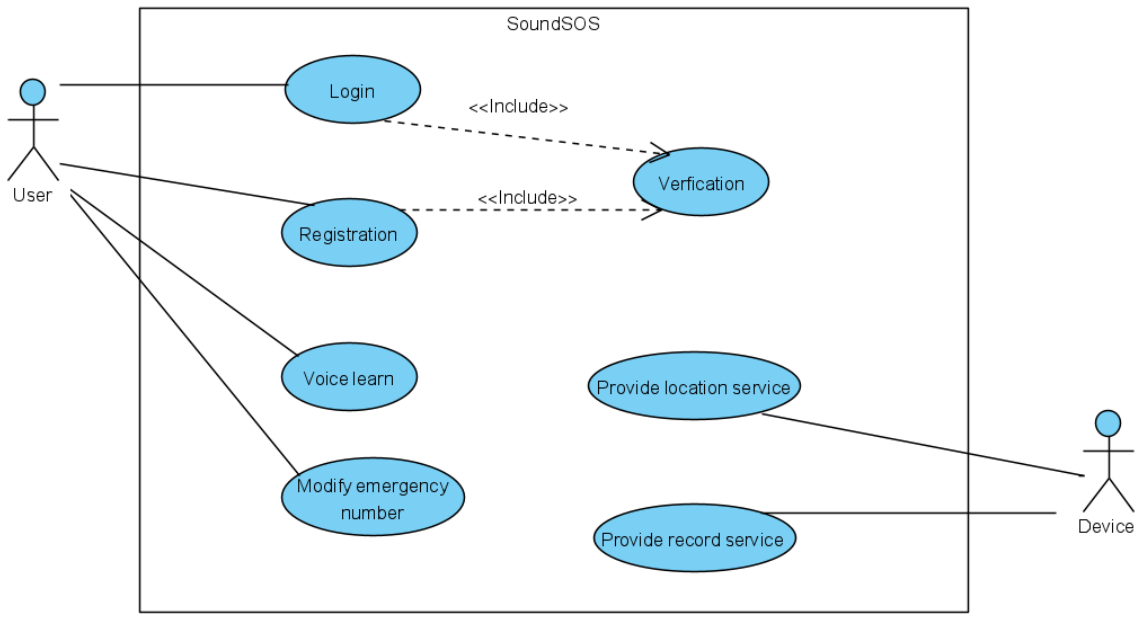



Figure 5.
Use-Case Diagram

Use Case Description:

Login - The user will fill username and the password fields to login into the application.

Registration - The user can register to the system for the login.

Verification- The system will check the details and present relevant output messages.
Voice Learned - The user will learn the system with his voice.

Modify emergency number - The user will have the option to modify emergency number if needed.

Provide location service - The user's phone will provide access to location service.

Provide record service -The user's phone will provide access to record service.

## Activity Diagram

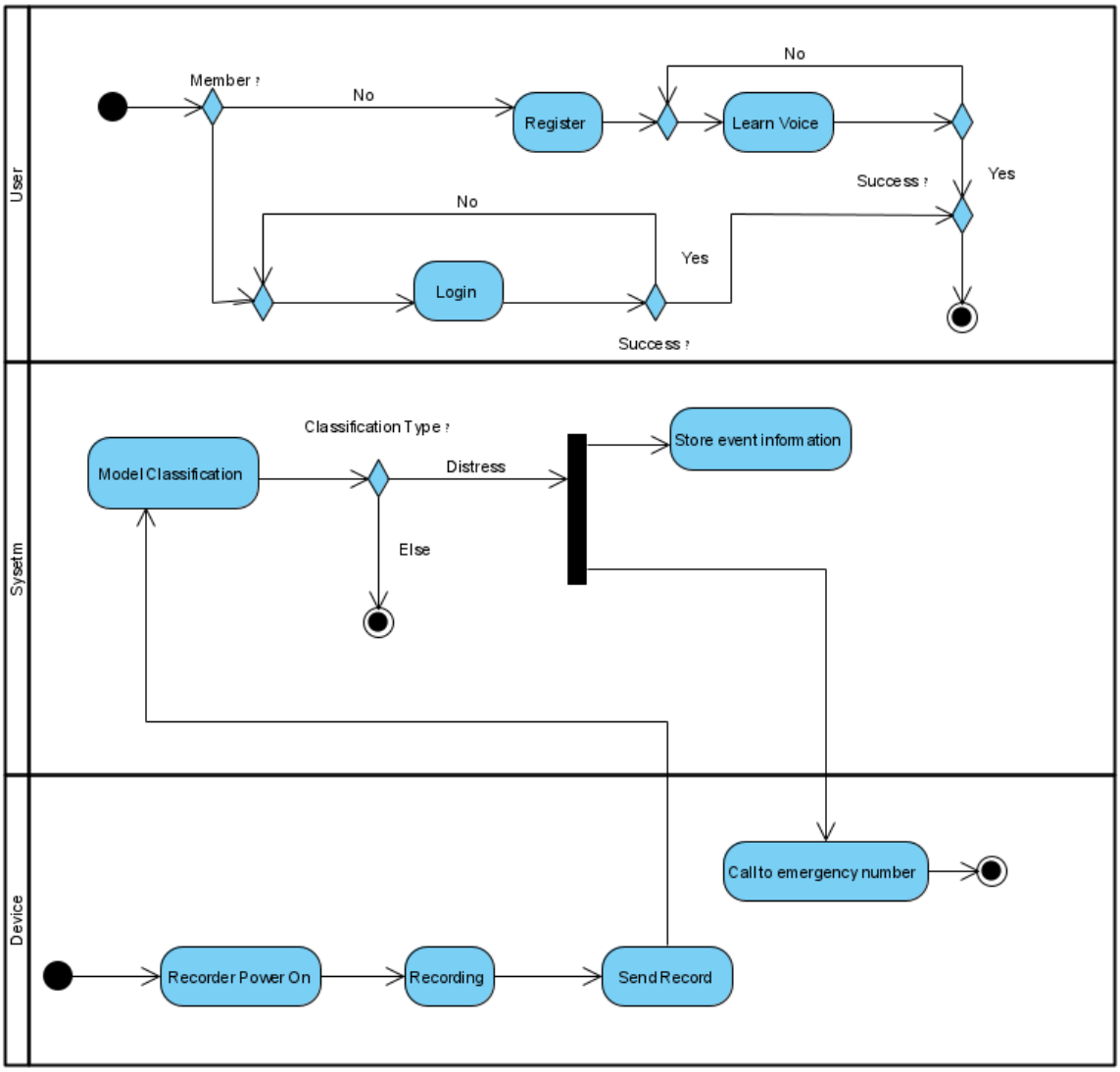


Figure 6.
Activity Diagram

Activity Description:

User – First, the user needs to register and then, he will need only to learn his voice.

Device – The device needs to provide an access to record service for recording and to provide the audio record to the system. Also, the device will provide access to an automated dial.

System – Our model will detect the emotion in the audio file and will perform the classification and will act accordingly.

## Class Diagram

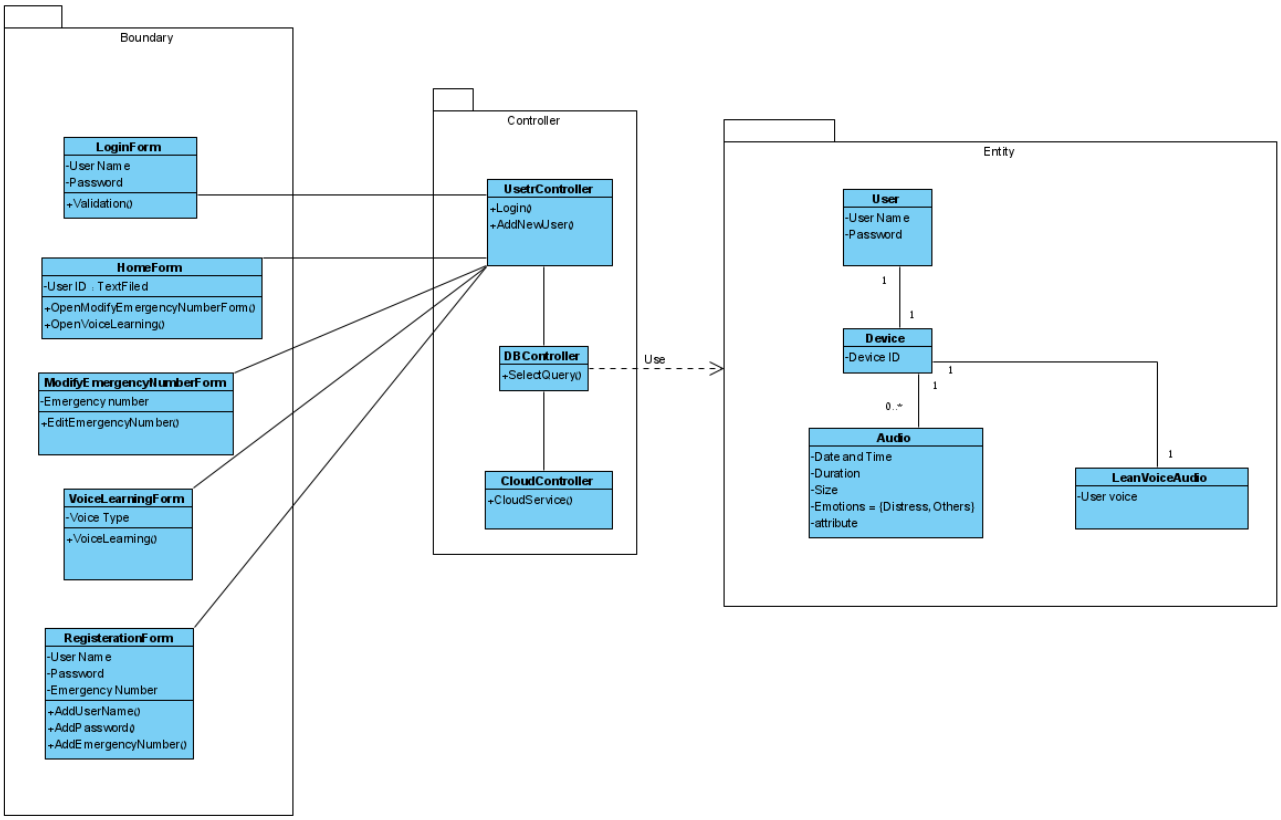
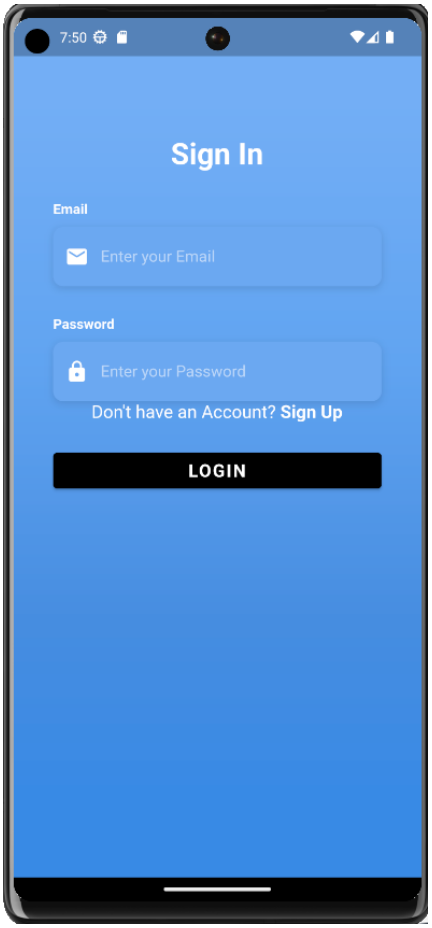
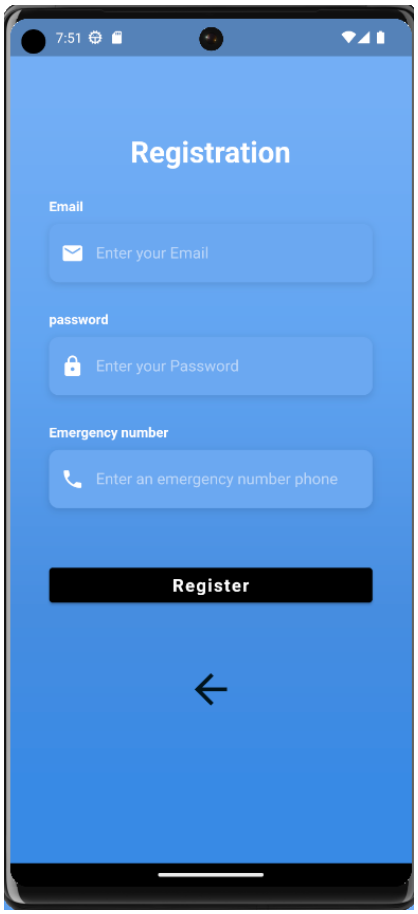
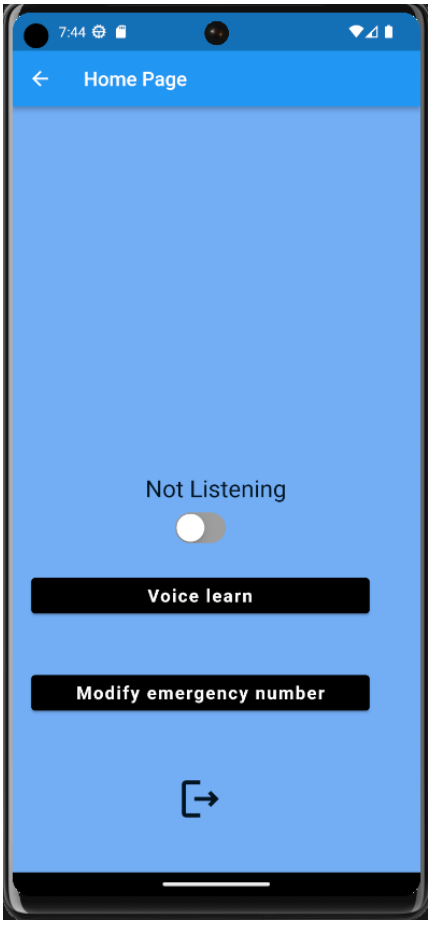


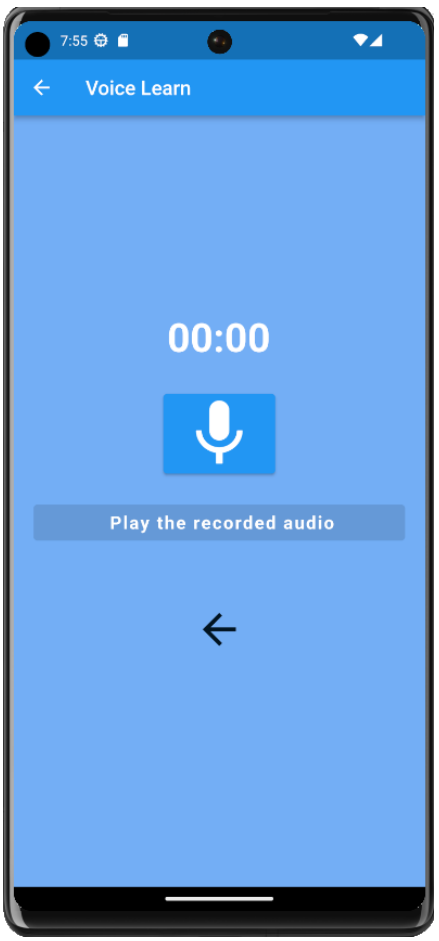

Figure 7.
Class Diagram

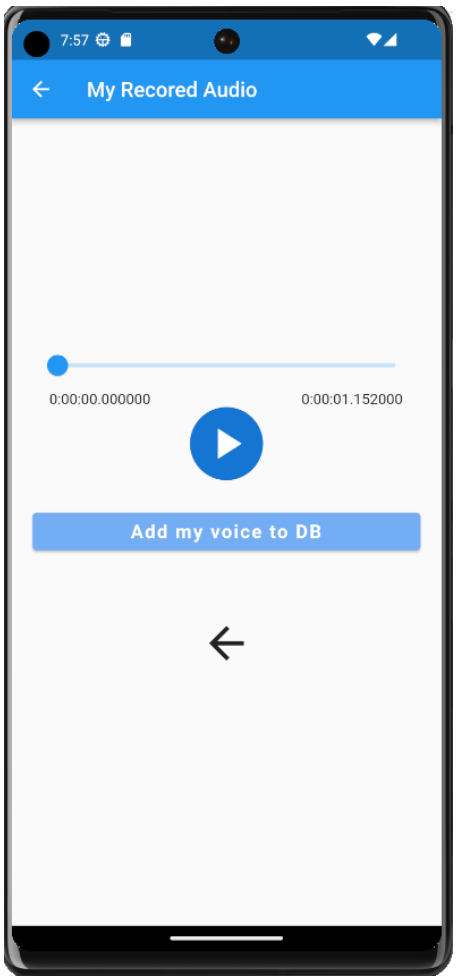
## User Interface

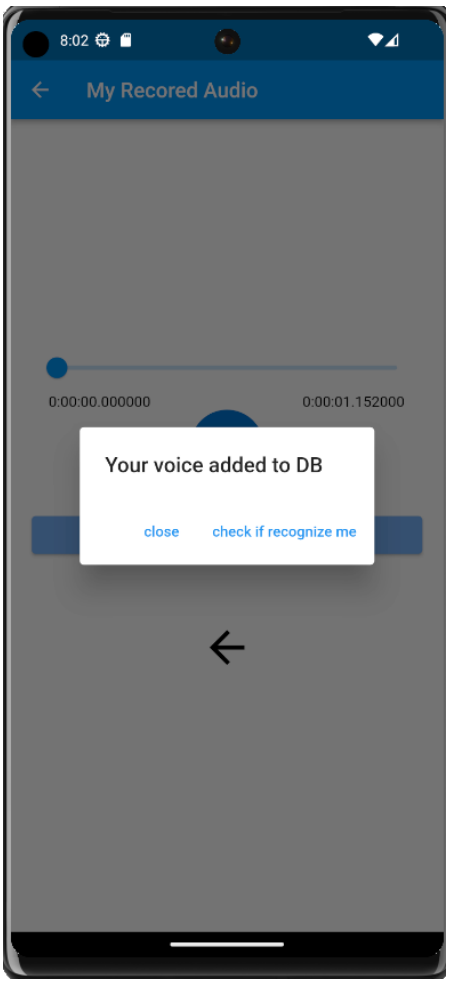


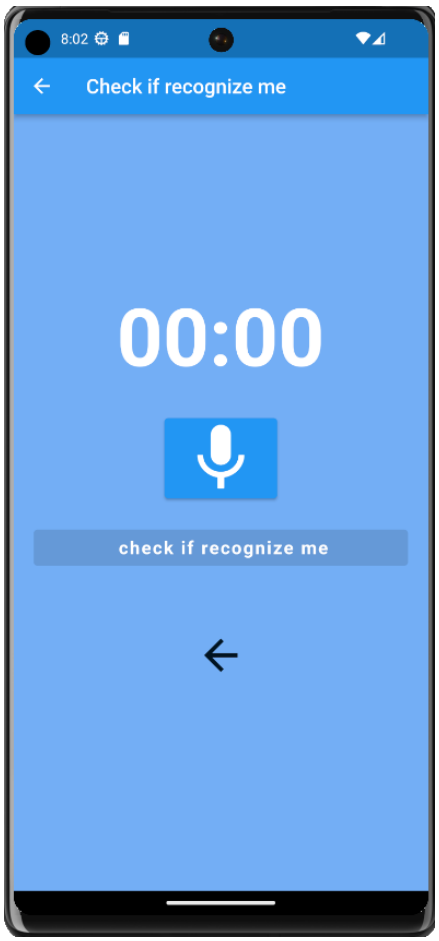



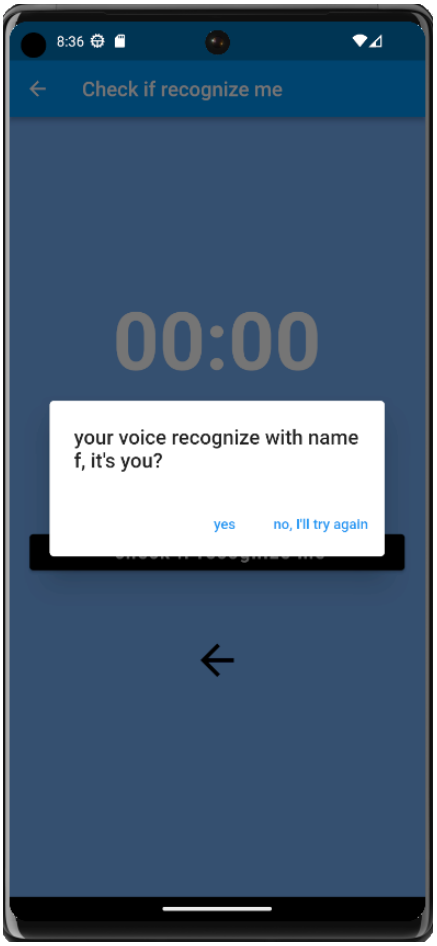




7:57
My Recored Audio
0:00:00.000000
0:00:01.152000
Add my voice to DB


8:02
My Recored Audio
0:00:00.000000
0:00:01.152000
Your voice added to DB
close
check if recognize me


8:02
Check if recognize me
00:00
check if recognize me


8:36
Check if recognize me
00:00
your voice recognize with name
f, it's you?
yes
no, I'll try again

# 5.Conclution

In our application, we have developed a novel feature for voice learning, which involves the use of a multi-layer CNN model. This model has been specifically designed to address two tasks: unique voice classification and emotion voice classification. By leveraging the power of convolutional neural networks, we aim to accurately classify voices based on their unique characteristics and emotional attributes.

During our experimentation and training phase, we achieved a validation accuracy of 70% with our existing model. This performance indicates that our model can make reasonably accurate predictions. However, we acknowledge that there is still room for improvement. One of the key factors that can significantly enhance the model's performance is the availability of more diverse and abundant training data.

To address this limitation, we have planned to further enhance our model's training process by collecting additional data from various background environments. By incorporating a wider range of voice samples into our training dataset, we expect our model to gain a better understanding of the variations and nuances present in different voices. This expanded dataset will enable the model to learn more robust and discriminative representations, ultimately leading to improved classification accuracy.

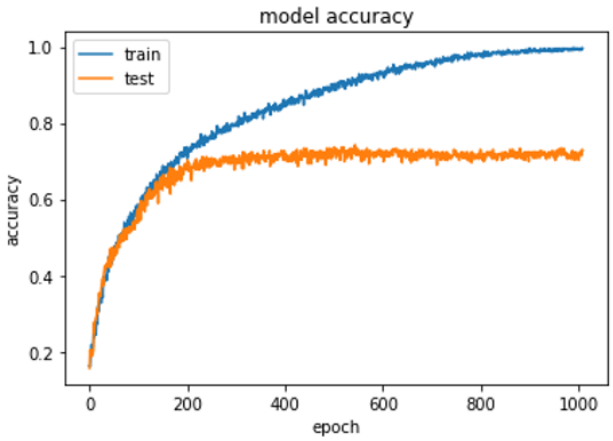



Figure 8.
Model accuracy

# 6. Verification and Evaluation

To ensure that all requirements have been completed, it is necessary to run all tests and verify that they all pass. This ensures that the system has been correctly implemented and that all requirements have been completed.

| Test | Screen | Scenario under check | Expected Result |
|---|---|---|---|
| 1 | Login | Correct Username & Correct Password and press sign in | Go to 'Home Page' screen |
| 2 | Login | Incorrect Password/Incorrect Username and press 'sign in' | Error message |
| 3 | Login | Press sign up | Go to Register |
| 4 | Register | Incorrect field type and press 'create account' button | Error message |
| 5 | Register | Correct field type and press 'create account' | Go to 'Home Page' screen |
| 6 | Home Page | Press 'Modify Emergency Number' button | Open popup 'Modify Emergency Number' |
| 7 | Home Page | Press 'Voice Learn' button | Go to 'Voice Learn' screen |
| 8 | Home Page | Enter 'ON' & Done level of voice learn | Turn on the listening |
| 9 | Home Page | Enter 'ON' & not Done level of voice learn | Open error popup |

| | | | |
|---|---|---|---|
| 10 | Home Page | Enter ‘OFF’ | Turn off the listening |
| 11 | Home Page | Press ‘Logout’ | Go to ‘Login screen’ |
| 12 | Modify emergency number (PopUp) | Incorrect field type(numbers only) | Block field |
| 13 | Modify emergency number (PopUp) | Press submit button | Go to ‘Home Page’ screen |
| 14 | Voice Learn | Press record logo button | Audio is created |
| 15 | Voice Learn | Press back logo button | Go to ‘Home Page’ screen |
| 16 | Voice Learn | Press "Play recorded audio" | Go to 'Audio Record' screen |
| 17 | Audio Record | Press on the play button | Play the recorded audio |
| 18 | Audio Record | Press on the "Add my voice to DB" button | Open popup with message from server |
| 19 | Audio Record(popup) | Press "check if recognize me" button | Go to "Check if recognize me" screen |
| 20 | Audio Record(popup) | Press "close" button | Go to " Audio Record " screen |
| 21 | Check if recognize me | Press record logo button | Audio is created |
| 22 | Check if recognize me | Press back logo button | Go to ‘Audio Record' screen |
| 23 | Check if recognize me | Press 'check if recognize me' | Open popup with message from server |

| | | | |
|---|---|---|---|
| 24 | Check if recognize me(Popup) | Press 'yes' | Go to ‘Home Page’ screen |
| 25 | Check if recognize me(Popup) | Press 'No, I'll try again' | Go to "Check if recognize me" screen |
| 26 | Server | Recognize distress audio | 1.Save the audio file<br>2. Automatic call to emergency number |
| 27 | Server | Not Recognize distress audio | Remove the audio file |
| 28 | Server | Connection with DB | Import and export data from/to the DB |
| 29 | Server | Connection with client | Send and receive information from/to client |

# 7.References:


1. Matheson, Jennifer L. "The Voice Transcription Technique: Use of Voice Recognition Software to Transcribe Digital Interview Data in Qualitative Research." *Qualitative Report* 12.4 (2007): 547-560.
2. Petrushin, Valery. "Emotion in speech: Recognition and application to call centers." *Proceedings of artificial neural networks in engineering*. Vol. 710. 1999.
3. Alu, D. A. S. C., Elteto Zoltan, and Ioan Cristian Stoica. "Voice based emotion recognition with convolutional neural networks for companion robots." *Science and Technology* 20.3 (2017): 222-240.
4. Johnson, William F., et al. "Recognition of emotion from vocal cues." *Archives of General Psychiatry* 43.3 (1986): 280-283.
5. Guay, Stéphane, Jane Goncalves, and Juliette Jarvis. "Verbal violence in the workplace according to victims' sex—a systematic review of the literature." *Aggression and violent behavior* 19.5 (2014): 572-578.
6. Jacobson, Neil S., et al. "Affect, verbal content, and psychophysiology in the arguments of couples with a violent husband." *Journal of consulting and clinical psychology* 62.5 (1994): 982.
7. Stewart, Catherine Carolyn, Debra Langan, and Stacey Hannem. "Victim experiences and perspectives on police responses to verbal violence in domestic settings." *Feminist Criminology* 8.4 (2013): 269-294.
8. Arriany, Aml A., and Mohamed S. Musbah. "Applying voice recognition technology for smart home networks." *2016 International Conference on Engineering & MIS (ICEMIS)*. IEEE, 2016.
9. Carbonell, Jaime G., Ryszard S. Michalski, and Tom M. Mitchell. "An overview of machine learning." *Machine learning* (1983): 3-23.
10. Nasteski, Vladimir. "An overview of the supervised machine learning methods." *Horizons. b* 4 (2017): 51-62.
11. Totakura, Varun, Bhargava Reddy Vuribindi, and E. Madhusudhana Reddy. "Improved Safety of Self-Driving Car using Voice Recognition through CNN." *IOP Conference Series: Materials Science and Engineering*. Vol. 1022. No. 1. IOP Publishing, 2021.
12. Zhao, Gansen, et al. "Modeling MongoDB with relational model." *2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies*. IEEE, 2013.
13. Abramova, Veronika, and Jorge Bernardino. "NoSQL databases: MongoDB vs cassandra." *Proceedings of the international C* conference on computer science and software engineering*. 2013.

14. Badrulhisham, Nur Alia Syahirah, and Nur Nabilah Abu Mangshor. "Emotion Recognition Using Convolutional Neural Network (CNN)." *Journal of Physics: Conference Series*. Vol. 1962. No. 1. IOP Publishing, 2021.

15. Chen, Shizhe, and Qin Jin. "Multi-modal dimensional emotion recognition using recurrent neural networks." *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. 2015.

16. Badshah, Abdul Malik, et al. "Speech emotion recognition from spectrograms with deep convolutional neural network." *2017 international conference on platform technology and service (PlatCon)*. IEEE, 2017.

17. Lukic, Y., Vogt, C., Dürr, O., & Stadelmann, T. (2016, September). Speaker identification and clustering using convolutional neural networks. In 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP) (pp. 1-6). IEEE.