Real-Time Facial Expression and Speech Emotion Recognition App Development on Mobile Phones using Cloud Computing

by

Humaid Saif Saeed Alasli Alshamsi

A dissertation submitted to the Department of Computer Engineering and Sciences
Florida Institute of Technology
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Computer Engineering

Melbourne, Florida
November 2018
We the undersigned committee hereby approve the attached thesis, “Real Time Facial Expression and Speech Emotion Recognition App Development on Mobile Phones using cloud Computing” by Humaid Saif Saeed Alshamsi.

Veton Kepuska, Ph.D.
Associate Professor, Department of Computer Engineering and Science

Samuel Kozaitis, Ph.D.
Professor, Department of Computer Engineering and Science

Josko Zec, Ph.D.
Associate Professor, Department of Computer Engineering and Science

Maria Pozo De Fernandez, Ph.D.
Assistant Professor, Department of Biomedical and Chemical Engineering and Sciences

Philip Bernhard, Ph.D.
Department Head of Computer Engineering and Science
Emotion plays a vital role in humans’ daily lives. Understanding emotions and recognizing how to react to others’ feelings are fundamental to engaging in successful social interactions. Emotion recognition through facial expression and speech play a significant role in human communication. This subject is becoming important in academic research as new techniques such as emotion recognition from speech context inspire us to recognize how emotions are related to the content we are uttering.

The demand and importance of emotion recognition have highly increased in many applications in recent years, such as video games, human-computer interactions, cognitive computing, and affective computing. Recognizing emotion is achieved from many sources including text, speech, hand, and body gestures as well as facial expressions. Most of the emotion recognition methods only use one of the sources mentioned previously. Human emotions change almost every second and using a single way to process the emotion recognition may not reflect it correctly.

The motivation for this research is based on my desire to understand and evaluate emotions in multiple ways such as facial and speech expressions.

The topic of my dissertation is an examination of Real-Time facial expression and speech emotion recognition on a mobile phone using cloud computing. The proposed framework can recognize emotion from facial expression as well as speech in real time, that was embedded into an application that was developed for mobile phone.
There are three parts in the design of the system: the facial emotion recognizer, the speech emotion recognizer, and merging both systems; the combined facial expression and speech recognition that runs on a smartphone using Cloud Computing (the app. name called Emotii). The Emotii Facial Expression and Speech part uses the results from the facial expression recognition and speech emotion recognition. Then, a novel method is used to integrate the results, when a final decision of the emotion is given after the fusion of those features.

The application works in real-time on any mobile phone that has an android operating system and is capable of displaying correct emotion. The result is given as a percentage of all emotions such as neutral, happy, sad, angry, surprise, disgusted, and fear. The experiment results demonstrate that the emotional face and speech recognition on a mobile phone has been successful and it gives up to 97.26% correct results as measured from standard corpora: a. Cohn-Kanade (CK+), b. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).
# Table of Contents

Table of Contents .................................................................................................................. v
List of Figures ............................................................................................................................ viii
List of Tables ............................................................................................................................... xi
List of Abbreviations ................................................................................................................ xii
Dedication ................................................................................................................................. xiv

## Chapter 1: INTRODUCTION ......................................................................................... 1
  1.1. Introduction ......................................................................................................................... 1
  1.2. Goals and Objectives ........................................................................................................... 2
  1.3. Research Question ............................................................................................................. 3
  1.4. Emotions and Emotion Recognition ................................................................................... 4
  1.5. Emotion Models .................................................................................................................. 6
  1.6. The Basic Emotion Models ............................................................................................... 7
  1.7. The Dimensional Model ..................................................................................................... 11
  1.8. The Componential Appraisal Models ............................................................................... 13
  1.9. Outline of the Dissertation ............................................................................................... 14

## CHAPTER 2: FACIAL EXPRESSION RECOGNITION .................................................. 15
  2.1. Introduction ......................................................................................................................... 15
  2.2. Facial Expressions .............................................................................................................. 18
    2.2.1. Physiology of Facial Expressions ................................................................................ 19
    2.2.2. Facial Action Coding System .................................................................................... 21
  2.3. Related Works .................................................................................................................... 22
  2.4. Factors of Facial Emotion Recognition ............................................................................. 25
    2.4.1. Face Detection ............................................................................................................ 26
    2.4.2. Face Tracking ............................................................................................................. 27
    2.4.3. Feature Extraction ...................................................................................................... 28
    2.4.4. Emotion Classification .............................................................................................. 29
  2.5. The Case for Affective Research Using Technology ......................................................... 29
    2.5.1. The Current State of Technology: Machines with Emotions and Personalities ........ 30
    2.5.2. Facial Detection Technology ..................................................................................... 31
  2.6. Studies in Facial Expressions ............................................................................................ 32
    2.6.1. KDEF Database ......................................................................................................... 33
    2.6.2. JAFFE Database ....................................................................................................... 33
    2.6.3. Cohn-Kanade Database ............................................................................................ 34
  2.7. The Current Literature on Affective Dimensions & Cognition ........................................ 35
  2.8. Facial Action Coding System ............................................................................................ 36
  2.9. Facial Expression Technology and Smart Phones ............................................................... 40
List of Figures

Figure 1: High light six emotional faces.................................................................9
Figure 2: Plutchik's wheel of emotions [12]............................................................10
Figure 3: Distinctive and dimensional emotion model [18].................................12
Figure 4: Muscles of facial expressions, taken from [33].....................................20
Figure 5: Facial emotion recognition steps............................................................26
Figure 6: Male subject is showing all the seven basics from the KDFE dataset ....33
Figure 7: Examples of images from the Japanese Female Facial Expression database.................................................................34
Figure 8: An image sequence of a subject expressing 'Surprise' from the Cohn-Kanade Facial Expression Database.........................................................35
Figure 9: A demonstration of how the Facial Action Coding system identifies muscle groups. .................................................................37
Figure 10: Upper face Action Units and some combinations [106].......................38
Figure 11: Some lower face Action Units and some combinations [106]. ........38
Figure 12: Overview of a speech emotion recognition system steps. .................54
Figure 13: Representation of a signal as a waveform and as a spectrogram.......56
Figure 14: Emotion relevant acoustic properties shown for a neutral, a happy and a bored utterance of an actress taken from the Berlin Database of Emotional Speech [31]. The text spoken in each of the utterances was “Das will sie am Mittwoch abgeben.” (“She will hand it in on Wednesday.”). ................58
Figure 15: An example of decomposition into frequency scales by wavelet transformation.................................................................60
Figure 16: The frequency spectrum of an /a/ in neutral, happy and boring speech.61
Figure 17: The times of glottal pulses (blue lines) marked in the same speech signals as in Figure 14.................................................................................63
Figure 18: Analog signals of a recorded speech ...............................................67
Figure 19: Digitizing of speech signal. .................................................................68
Figure 20: PCA: a mapping of the original two dimensions onto the principal component axes.................................................................74
Figure 21: Difficulty of speech emotion recognition with different types of databases. ........................................................................77
Figure 22: Overview Structure of Real Time Facial Expression and Speech Emotion Recognition System.................................................................78
Figure 23: Overview structure of the real time emotional face recognition face on mobile phone system.................................................................79
Figure 24: Haar features [199]. ...........................................................................81
Figure 25: (a) Facial Landmarks and COG using CK+ dataset (b) Line mapping between COG and Facial Landmarks. (c) Face offset correction. ..............83
Figure 26: An optimal hyperplane that classifies the two classes by SVM classifier.

Figure 27: Examples of six basic facial expressions from the CK+ database Source: CK+database (© J. Cohn).

Figure 28: KDEF Facial Expression Dataset.

Figure 29: JAFFE Facial Expression Dataset.

Figure 30: Example from CK+ dataset used to Extract the feature using MATLAB.

Figure 31: Histogram of Facial Landmarks descriptors using CK+ dataset.

Figure 32: Overview of testing facial expression through the app.

Figure 33: Column visualization of three detected emotions of “Surprise,” “Sadness” and “Happiness.”

Figure 34: Line with visualization marks of three detected emotions of “Anger,” “Natural” and “Happiness.”

Figure 35: Pie visualization of three detected emotions of “Anger,” “Disgust,” “Fear,” “Surprise,” “Sadness”, and “Happy”.

Figure 36: Typical steps for speech emotion recognition [155].

Figure 37: Speech signal processing.

Figure 38: Block diagram of the computation steps of MFCC.

Figure 39: Support Vector Machine structure.

Figure 40: Example from RAVDESS dataset used to Extract the feature of Sadness using MATLAB.

Figure 41: Example from RAVDESS dataset used to Extract the feature of Happiness using MATLAB.

Figure 42: Example from RAVDESS dataset used to Extract the feature of Fear using MATLAB.

Figure 43: Example from RAVDESS dataset used to Extract the feature of Disgust using MATLAB.

Figure 44: Example from RAVDESS dataset used to Extract the feature of Anger using MATLAB.

Figure 45: Anger Speech Emotion through the app.

Figure 46: Disgust Speech Emotion through the app.

Figure 47: Happiness Speech Emotion through the app.

Figure 48: Sadness Speech Emotion through the app.

Figure 49: Neutral Speech Emotion through the app.

Figure 50: Surprise Speech Emotion through the app.

Figure 51: Block diagram of Audio-Visual Emotion Recognition using Feature-Level Fusion.

Figure 52: Block diagram of Feature-Level Fusion.

Figure 53: Support Vector Machine structure.

Figure 54: Overview of testing fusion through the app.
Figure 55: The sample face expression images from the Cohn-Kanade database
(©Jeffrey Cohn) [68].

Figure 56: The comparison of multi-sensory emotion recognition.
List of Tables

Table 1: Facial Expression Description of six Basic Emotions [21].........................16
Table 2: Description of several AUs together with the muscle name [36]......................21
Table 3: Basic AUs connected with emotions [37].................................................22
Table 4: Mobile application detail [44].................................................................44
Table 5: Variations of acoustic variables observed in emotional expressions compared to neutral speech. Synopsis of Murray and Arnott [122]’s summary table..........49
Table 6: Different emotion units of speech............................................................69
Table 7: A summary of LLDs/short-term features....................................................71
Table 8: Confusion matrix SVM classification and number in the 1th database of Cohn and Kanade (CK+) dataset.................................................................128
Table 9: Confusion matrix SVM classification and a number of images in the 2nd database of Japanese Female Facial Expression (JAFFE) dataset.................................128
Table 10: Confusion matrix SVM classification and number of images in the 3rd database of The Karolinska Directed Emotional Faces (KDEF) dataset.............129
Table 11: Confusion matrix using SVM classification for CK+ dataset.......................130
Table 12: Confusion matrix using SVM classification for JAFFE dataset......................130
Table 13: Confusion matrix using SVM classification for KDEF dataset.....................131
Table 14: Total number of images collected for Facial Expression Recognition study. ..........................................................131
Table 15: Confusion matrix SVM classification and number of wave files in the 1th dataset of Ryerson Audio-Visual Database of Emotional Speech and Song Dataset (RAVDESS).............................................................................................................134
Table 16: Confusion matrix SVM classification and number of wave files in the 2th dataset of Ryerson Multimedia Research Lab (RML).................................................134
Table 17: Confusion matrix SVM classification and number of wave files in the 3th dataset of Surrey Audio-Visual Expressed Emotion (SAVEE).........................................................135
Table 18: Confusion matrix using SVM classification for RAVDESS dataset..............136
Table 19: Confusion matrix using SVM classification for RML dataset.......................136
Table 20: Confusion matrix using SVM classification for SAVEE dataset..................137
Table 21: Total number of wave files collected for speech emotion study..................137
Table 22: Confusion matrix using SVM classification for RAVDESS dataset..............138
Table 23: Confusion matrix using SVM classification for RML dataset.....................139
Table 24: Confusion matrix using SVM classification for SAVEE dataset..................139
Table 25: Confusion matrix using SVM classification for RAVDESS dataset.............140
Table 26: Confusion matrix using SVM classification for RML dataset.....................140
Table 27: Confusion matrix using SVM classification for SAVEE dataset..................140
Table 28: Total number of wave files collected for each part of the study..................141
Table 29: The detailed accuracy of emotion recognition..........................................142
Table 30: Accuracy Comparison Between the Systems.............................................145
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFER</td>
<td>Automatic Facial Expression Recognition</td>
</tr>
<tr>
<td>ASER</td>
<td>Automatic Speech Emotion Recognition</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>FER</td>
<td>Facial Expression Recognition</td>
</tr>
<tr>
<td>SER</td>
<td>Speech Emotional Recognition</td>
</tr>
<tr>
<td>CK+</td>
<td>Cohn-Kanade Dataset</td>
</tr>
<tr>
<td>JAFFE</td>
<td>Japanese Female Facial Expression Dataset</td>
</tr>
<tr>
<td>KDEF</td>
<td>Karolinska Directed Emotional Faces Dataset</td>
</tr>
<tr>
<td>SAVEE</td>
<td>Surrey Audio Visual Expressed Emotion Speech Dataset</td>
</tr>
<tr>
<td>RML</td>
<td>Ryerson Multimedia Research Laboratory Dataset</td>
</tr>
<tr>
<td>RAVDESS</td>
<td>Ryerson Audio-Visual Database of Emotional Speech and Song</td>
</tr>
<tr>
<td>FACS</td>
<td>Facial Action Coding System</td>
</tr>
<tr>
<td>COG</td>
<td>Center of Gravity</td>
</tr>
<tr>
<td>RBF</td>
<td>Gaussian radial basis function</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Predictive Coding</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel-Frequency Cepstral Coefficient</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transformation</td>
</tr>
<tr>
<td>DCT</td>
<td>Discrete Cosine Transformation</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transformation</td>
</tr>
<tr>
<td>AUs</td>
<td>Action Units</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>OF</td>
<td>Optical Flow</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Predictive Coding</td>
</tr>
</tbody>
</table>
Acknowledgment

My sincere thanks and deep gratitude to Almighty ALLAH; Him without whose help and guidance the completion of this work would not have been possible.

I would like to express my gratitude to my supervisor, Dr. Veton Kepuska, who has helped me to complete this dissertation through his sincere advice, unfailing assistance, guidance, patience, continuous encouragement, and understanding throughout the course of this study.

I would like to thank my doctoral committee, Dr. Samuel Kozaitis, Dr. Josko Zec, and Mussafar Shaikh, for providing a very conducive research atmosphere and thoughtful comments.

Additional thanks are due to all of my colleagues and friends of their help, support, and encouragement during my studies.

Last, but not least, I would like to express my sincere and heartful gratitude to my parents for their prayers and patience during my studies. I know that whatever I say, I shall not qualify and compensate them. Their presence with me here in the United States has given me much strength and encouragement to move forward and to accomplish what they and I have dreamed of.

A very special thanks goes to my brothers and sisters. Without their selfless support and encouragement, I could not complete the research and dissertation and my life in general.
Dedication

I hereby declare that the work presented in this thesis is based on the original work done by me under the supervision of Dr. Veton Kepuska, Associate Professor, in the Department of Electrical and Computer Engineering, Florida Institute of Technology, Melbourne, Florida (USA) and that no part thereof has been presented for the award of any other degree.

Humaid Saif Alshamsi
Department of Computer Engineering and Science
School of Engineering
Florida Institute of Technology
Chapter 1: INTRODUCTION

1.1. Introduction

Computers play an essential role in our lives. Traditional interfaces of computers like the keyboard and the mouse are not able to meet today’s requirements in different human activities. We need a more natural interaction between computers and humans. Currently human-computer interaction is far less natural than the human interaction with each other. It is not possible yet to do face-to-face communication through human-computer interaction. To improve the neutrality, computers should be able to support the way humans communicate with each other. The essential requirement is that they should have the ability to acquire and show emotions. The computer should be able to recognize the emotions of the human counterpart to have a more natural interaction that mimics the interaction between humans. According to many theories of emotion like Meyers (2004) [1], in humans, emotion is one of the most complicated, physical, and psychological behaviorss [1]. In daily social life, knowing the emotional feeling of the communication partner is critical and intuitive. However, when it comes to the computer, it is much harder for it to determine the emotion of its counterpart. In human communication, there are many aspects such as speaking, body gestures, eye contact, and facial expressions that help us in understanding each other. The basic units of verbal communication are words in speech, and the basic units of the nonverbal communication include the facial expression, body movements, and gestures as well as the eye contact. Sometimes, speech itself is enough for communication; a phone call is a good example. Nonverbal communication units are also important, especially in face-to-face communication. The emotions reflected in facial expressions are far more intense than words. To achieve a more natural human-computer interaction, the computers need to have the ability to recognize the emotion from speech, facial expression, or
body gestures. In this thesis we propose Android application called an Emotii (it is Facial Expression and Speech Recognition system) framework to recognize emotion from speech and facial expression. Experiments show that Emotii Facial Expression and Speech Recognition can help a smartphone recognize human emotions more accurately. The proposed Emotii Facial Expression and Speech Recognition on a smartphone will help to improve the human-computer interaction.

1.2. Goals and Objectives

The goal of this research is to design an application that runs on the Android platform for analyzing the emotional facial expression and speech of a person in real time using a Cloud Computing Server. We will demonstrate that the system can recognize seven different basic emotions: happy, sad, surprised, fear, angry, disgust and neutral with the highest reported accuracy. By using the facial expression and speech analysis, the detection of facial muscle actions and vocal voice, the system has reached a stage that makes it practical to integrate its analysis with a video camera. This research aims to introduce an Android application that uses a camera and microphone for mobile phone in real time to detect the face and speech and then display the emotions of the face and speech. The user can see a real-time emotional expression of face and speech that will be displayed automatically on a phone screen. The application will be designed to enable the display of seven different emotions on the top side of the phone screen. The user has three options to choose from this app such as (1- Facial expression recognition, 2-speech emotion recognition, 3-combined facial and speech emotion recognition) by choosing one of these options and the video/audio or both it will detect the right expression such as (e.g., happiness, sadness or surprised) from face/speech or both and then display recognized emotion on the screen. Once the operation is complete, the application will display the result to the user.
This research also aims to investigate the methodology for detection and recognition of emotional facial expression and speech displayed. In a real situation, facial expressions and speech have varying intensities, and many practical applications would benefit from systems that can compare the intensities of two or more expressions. Main computing tasks will be implemented in the Cloud Server, and a free application on a mobile phone can be downloaded by a user. A database and or corpora can be built through the smartphone application. The system will be used by many cloud-based Internet applications. The project involves machine learning algorithms and the use of computer vision applications.

**The general objectives of the project are:**

- Develop a system of how expressions of different intensities could be ranked, and then apply suitable machine learning algorithms to predict where a new unseen image/audio would fit in this ranking.
- Implement the system in Android platform using Java language.
- Create a full application that uses speech recognition, speech processing, image processing and machine learning on Android OS.
- Collect a database of facial expression images and speech emotion audio that will be used to display a Real-Time emotional face and speech.

Develop an application that gives users the ability to test their own facial and speech emotions.

**1.3. Research Question**

- Why, and with what impact on Real-Time emotional face and speech recognition system will impact cloud computing server in future phone application?
- How well does the AFER (Automatic Facial Expression Recognition) method extract the facial expressions of the adults /children in the Cohn-Kanade
(CK+), Japanese Female Facial Expression (JAFFE), and Karolinska Directed Emotional Faces (KDEF) datasets?

- How well does the ASER (Automatic Speech Emotion Recognition) method extract the speech emotion of the adults/children?
- When integrating the recognition of facial emotion and speech, how the system performance compared to the individual?

**Research Objectives**

a. To describe and provide evidence of work by simulation and provide necessary calculations.

b. To select one algorithm that will be used and make further investigation.

c. To investigate the relevant literature for considerations of why this work is now taking place, including considerations of application development.

d. To understand the primary concept drivers contributing to these decisions to enter this area of work.

**1.4. Emotions and Emotion Recognition**

From a psychological perspective, emotions are defined as the complicated state of a feeling which reflects the mental and physical changes witnessed by an individual regarding their behavior and thoughts. Emotions are derived from speech, gestures, voices, as well as general facial expressions that play a significant role in our daily lives. The facial expression of smiling suggests that the person may be happy at that particular moment, but if someone is frowning, then this may indicate that they are upset or angry.

Individuals use a significant number of emotions on a daily bases, thus the difficulty in quantifying the number of emotions that a human can express. In the field of psychology, it is widely accepted that human emotions fall into 6 primary categories which consist of surprise, disgust, fear, sadness, anger, and happiness. Humans also
express their various moods and feelings by displaying diverse facial expressions and gestures when they communicate. Some emotional expressions fit into more than one category.

Human-computer interactions such as in video games, understanding and recognizing emotions will play a significant role in the evolution of the technology to create an enhanced experience for the player.

Therefore, the recognition of emotions and its application has gained a considerable amount of popularity both from a research and industry perspective. The observation and interpretation of human emotions is a challenging task for computers since emotions in themselves are complex.

The recognition of emotions can be conducted in numerous ways such as gestures, voice, facial expressions whereby the recognition of facial emotions and speech emotions receives an increasing interest in the designing of the human-machine communication system. Oral communication is a vital source of information on emotion recognition when people interact with others. In certain circumstances it is more important how we said it then what we said. Facial expressions are the most visible type of emotion communication, and therefore it is considered to be easily controlled by the person in the various social environments. As a result, the recognition of emotions in one single way may not accurately reflect the individual’s emotional state.

To be able to overcome the limitations of singular emotion recognition, the multimodal emotion recognition was theoretically proposed. Schuller, Lang, and Rigoll (2002) [2] have examined the multimodal emotion recognition by conducting an analysis on the user’s speech and haptic [2] interaction on a touchscreen or using a mouse. In their analysis, they used both common speech features in addition to semantic and intension-based features. Multimodal emotional recognition derived
from facial expressions, speech and body gestures was initially presented by Caridakis et al. (2007) [3] in which the Bayesian Classifier was used to classify the features which were extracted. The usage of the feature level and decision level method led to an increase in the accuracy of the results by more than 10% in comparison to the results concluded from gesture emotion recognition. Soleymani, Pantic, and Pun (2012) [4] had proposed a user independent emotion recognition method through the usage of electroencephalogram signals (EEG), eye gaze data and the bodily responses to videos [4] with the purpose of recovering useful tags for videos. In their research study, Soleymani, Pantic, and Pun (2012) [4] adopted a 2D emotion model. The limitation of this study was the utilization of only a small video data set during the work.

1.5. Emotion Models

There are numerous challenges in the research relating to the recognition of emotions including what type of emotions ought to be conceived in order to recognize where these emotions are observed, given the information about emotions can be seen in conversations, speeches, body gestures, and facial expressions. A human is an expert in using emotions. We are also able to name the emotion that we have expressed to others and to identify the emotional state of the people with whom we communicate. On the other hand, the description of emotions by computers is difficult. The classification of emotions according to specific rules (e.g., real-time vs. not real-time) is even more difficult, it is for this reason that emotion modals remain critical. Psychologists have completed numerous detailed studies on emotions and have created proposals of multiple modals and theories that can be used to describe emotions. In general, emotions can be categorized into three main classifications: the basic emotion model, the componential appraisal model and the dimensional model [5].
1.6. The Basic Emotion Models

The Basic Emotion Models is based on the research of the emotion theory proposed by Darwin [6] which was interpreted by Tomkins [7][8]. Ekman was able to represent the theoretical proposal of the basic emotion model in Ekman [9]. According to Ekman’s work Ekman [9] (1992), upon the judgment of static images with human facial expressions, it can be asserted that six basic emotions are universally recognized. The six emotions are disgust, anger, fear, surprise, sadness, and happiness. The higher-level emotion recognition is not in the scope of this dissertation given that the majority of recent emotion recognition researches concentrate on the basic emotion model such as Ekman’s six basic emotions theory.

Emotional states and emotions have different definitions. The concept of emotions is based on feelings which are considered to be a subjective experience of an emotion. On the other hand, an emotional state can be measured based on the physiological changes that occur throughout the body but mainly in the face as a response to a specific emotion or emotions. These physical changes extend to the changes that can occur in blood pressure and hormones. However, facial expressions are considered to be the most significant channel used to express the emotional state of an individual.

Ekman is one of the researchers that tried to identify the relationship between emotions and facial expressions. According to Ekman’s theory, the emotions which are expressed at the same time accompanied by facial expressions pose as the social signals which can assist with building interpersonal communication skills [11]. Ekman distinguishes the six basic emotions which are surprise, disgust, anger, sadness, and happiness. These emotions correspond to the universally accepted facial expressions that are used in diverse cultures and form the basic building blocks of more complicated emotions. Collectively with Friesen, an accurate description of the facial features which correspond to emotion was formed [12].
Anger: Anger tends to be provoked by some action such as feelings of hurt, physical threats or frustration. The intensity of anger can vary from being irritated to furious. Those who experience anger may become violent. Anger can be expressed on the face with the eyebrows being lowered and drawn towards each other, wide eyes and the staring in one direction. The lips can be either pressed together or squarely shaped and parted. The salient facial indicator is in the eyes, the wide opening of the eyes and the drawn eyebrows (see Figure 1).

Disgust: Disgust tends to involve the feelings of aversion regarding tasting, smelling, touching or looking at something. The response for mild disgust would be to dislike something, to turn away whereas extreme disgust can even involve vomiting. Disgust can manifest itself facially with the rise of the upper lip; the lowered eyebrows and the wrinkled nose (see Figure 1).

Fear: Fear can occur when a person expects a particular event which can psychologically or physically harm them. It can range from being apprehensive to being terrified. Fear is considered to be an intense form and is considered the most traumatic type of emotion. It is characterized by the eyebrows slightly raised and drawn together with the lips stretched. The eyes tend to be opened with a tense lower lip (see Figure 1).

Happiness: Happiness is the most positive emotion that people experience with the states of relief, pleasure, and excitement. It is mostly expressed using the mouth in the shape of a smile. In the case of extreme happiness, the eyes are narrowed whereby crow feet wrinkles appear surrounding the outer corners of the eyes (see Figure 1).

Sadness: although commonly thought to be the opposite of happiness, it is not. It is the feeling that is caused by defeat, disappointment or loss. Sadness can last for extensive amounts of times from hours to even days. It can vary from being gloomy
to experiencing deep mourning. Sadness is indicated by the inside corners of the eyebrows being drawn and raised together, and the edges of the lips pull down (see Figure 1).

**Surprise:** an unexpected occurrence induces surprise. It is generally short time in duration, and since a person can have time to think about a surprising event, then they are not considered to be surprised. It is indicated with widened open eyes and dropping of the jaws (see Figure 1).

![Figure 1: High light six emotional faces.](image-url)
The six basic emotions exist in every culture, and they are intrinsic. Alongside these six fundamental emotions, there are numerous different emotions communicated by individuals around the World consistently. Those feelings are culture subordinate; they are learned as opposed to inborn and are combined from the fundamental feelings. They frame families emerging out of the basic emotions [12]. Every individual from emotion family shares certain qualities, for example, the same physiological exercises or encouraging occasions.

Plutchik agreed with Ekman's hypothesis and built the "wheel of emotions" [12]. In Figure 2, the Plutchik's wheel, eight emotions are organized in inverse sets (joy and sadness; anger and fear; disgust and acceptance; surprise and anticipation) with the quality of the emotions portrayed in unmistakable hues. As per Plutchik's examination, individuals cannot experience opposite emotions simultaneously. Complex feelings, which could emerge from a social condition or relationship with essential feeling, can be shaped by merely adjusting some fundamental feelings. Even though a few scientists have proposed an alternate number of fundamental feelings which can go from 2 to 18 [13] [14], Ekman's hypothesis of the six basic emotions is the most satisfactory one in the exploration of emotion recognition.

Figure 2: Plutchik's wheel of emotions [12].
The theory of basic emotions is not difficult to acknowledge, and its legitimacy is addressed. [13] The authors contended that the perspective of existing basic emotions (for instance, happiness and fear) could manufacture or clarify every single other emotion (disdain) is addressed, and the outflow of emotions is merely the same as the emotions themselves. For instance, special outward appearances that are perceived far and wide and appear to be general are not connected to emotions, instead to specific conditions that likewise inspire emotions. Moreover, they are usually insufficient to depict the emotions of everyday life by the blending of fundamental feelings. A genuine feeling might be amongst euphoria and shock, shadowed by expectation, and the power of the considerable number of parts is likewise essential. Ekman additionally revealed that there is some perplexity from the judicial investigation of the six basic emotions. For instance, anger and disgust, as well as fear and surprise, are commonly confused. The surprise is additionally mistaken for the emotions of intrigue.

1.7. The Dimensional Model

Emotions can also be reflected in a dimensional system where certain factors can map emotions. As indicated by this approach, the emotional states are not autonomous from each other. Considerations for a two-dimensional space are generally valence and arousal, as third-measurement energy or control is usually included. The valence measurement often speaks to the positive or negative level of the emotion, and the range is from awkward emotions to agreeable sentiments. The excitement measurement speaks to how energized the feeling is, and it ranges from low to high. The energy or control measurement speaks to the level of the energy or control over the feeling. These factors empower a more precise depiction of feelings, since different parts of feelings are utilized all the time in a constant scope of qualities. Fundamental feelings can be spoken in a dimensional model as a point or a zone. As appeared in Figure 3, as indicated by the distinctions of valence and
excitement, emotion can be spoken in the two-dimensional space. For instance, happiness can be inspired by wonderful dusk or a grinning infant (the yellow rectangle) in the excitement and valence measurements. There are a few examinations of utilizing dimensional models for emotion recognition [15] [16], and typically just two-dimensional models (valence-arousal model) are being used.

The benefit of the dimensional model is that it is exceptionally natural to speak to feelings on some constant scale. Also, a portion of the essential feelings proposed by Ekman, for example, sadness and happiness are not difficult to perceive in the dimensional models. Some different emotions, for example, anger and disgust are challenging to recognize, and a portion of the emotion cannot be depicted. Take Figure 3 for instance; surprise is absent in the two-dimensional model.

![Figure 3](image-url)

**Figure 3**: Distinctive and dimensional emotion model [18].
1.8. **The Componential Appraisal Models**

The componential appraisal model demonstrates the concerns relating to recognizing emotion as indicated by the translation of occasions which cause emotions outside. This model can be viewed as an augmentation of the dimensional model proposed by Scherer [19]. It sees feeling as a dynamic scene including a procedure of persistent change in every single relative segment, for example, cognition, motivation, physiological reactions, motor expressions, and feelings.

In the componential appraisal model, emotions are characterized as unpredictable, multi-componential, dynamic process, and there is no restriction on the numbers and the dimensional space of emotions. This model concentrates on the changing of enthusiastic states, predicts, and gives multiple versions of evaluation designs. The dissipation of feeling separation of this model enables researchers to identify individual differences and emotional disorders [19]. Nonetheless, the estimation of emotional states changing in this model is unpredictable and unsophisticated. In this manner, it is as yet an open region for emotion recognition with the appraisal model.

The basic emotion model discussed in this segment can essentially represent a significant portion of emotions and portray them effectively for psychological purposes. The six basic emotions proposed by Ekman are broadly utilized as a part of programmed emotion recognition because a higher number of measurements given by different models are not reliable enough to evaluate. In our dissertation, the six basic emotion model is selected in the design of an efficient emotion recognition system.
1.9. Outline of the Dissertation

The remainder of this dissertation is organized as follows:

Chapter 2 includes the background and related works about the facial expression recognition. We discussed in detail the general procedures for face detection, face tracking, feature extraction, and classification.

Chapter 3 introduces the background, related works on speech emotion recognition, and emotional speech databases.

Chapter 4 presents our framework of Merging Facial Expression and Speech Emotion Recognition using cloud computing that runs on a smartphone. There are three parts that are concerned with the speech emotion recognizer, facial expression recognizer, and both systems working together.

Chapter 5 is about the testing that we have done on the proposed system including choosing the base model of the speech recognizer, testing of speech and facial emotion recognition, results of merging facial with speech emotion recognition, as well as the testing on a static image for facial expression recognition.

Finally, in Chapter 6, some conclusion and future work are summarized.
CHAPTER 2: FACIAL EXPRESSION RECOGNITION

2.1. Introduction

Facial expressions are essential for effective social communication as they hold direct critical clues about emotional expressions. Facial movements play different roles in the interactions that occur between individuals. Some visual facial changes can be used to augment speech by enhancing emotions. Facial expressions remain actuated for a short while when the emotion is released. Therefore, the detection of facial expressions is an automatic method used for the recognition of emotions.

Ekman’s cross-cultural study [10] indicates that some emotional expressions are shared by all human beings irrespective of their region or race. Some of these emotions are surprise, disgust, fear, anger, sadness, and happiness. Each one of these emotions generally corresponds to a particular facial expression, in addition to other culturally diverse emotional expressions. Table 1 describes each of the facial expressions that correspond to the six basic emotions.

Facial expression analysis is considered to be a vital indicator of human affective state regarding emotion recognition systems. Numerous research fields such as psychology, computer vision and machine learning are involved in the multidisciplinary enterprise of automatic emotion recognition based on facial expressions. The facial muscle action detection and facial affect recognition are the two main current fields in facial emotion recognition. These two streams derive from two critical facial expression analysis methods found in psychological studies which are the sign judgment method and the message judgment method. The sign judgment method is generally used to describe the behavior displayed on the outside while the message judgment method is used to gain an understanding of what underlies the displayed facial expression.
Table 1: Facial Expression Description of six Basic Emotions [21].

<table>
<thead>
<tr>
<th>No</th>
<th>Emotion Name</th>
<th>Description of Facial expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Happiness</td>
<td>The eyebrows are relaxed. The mouth is open and the mouth corners upturned.</td>
</tr>
<tr>
<td>2</td>
<td>Sad</td>
<td>The inner eyebrows are bent upward. The eyes are slightly closed. The mouth is usually relaxed.</td>
</tr>
<tr>
<td>3</td>
<td>Fear</td>
<td>The eyebrows are raised and pulled together. The inner eyebrows are bent upward. The eyes are open and tense.</td>
</tr>
<tr>
<td>4</td>
<td>Anger</td>
<td>The inner eyebrows pulled downward and together. The eyes are wide open. The lips are tightly closed or opened to expose the teeth.</td>
</tr>
<tr>
<td>5</td>
<td>Surprise</td>
<td>The eyebrows raised. The upper eyelids and the eyes are wide open. The mouth opened.</td>
</tr>
<tr>
<td>6</td>
<td>Disgust</td>
<td>The eyebrows and eyelids are relaxed. The upper lip is raised and curled, often asymmetrically.</td>
</tr>
</tbody>
</table>

The FACS (Facial Action Coding System) that was developed by Friesan and Ekman is a widely used tool based on a vision which can code human expression movements based on their visual aspect displayed on the face which belongs to the sign judgment method.

The FACS allows the analysis of a facial expression to occur through the standardized coding of specific changes in facial movements with regards to Action Units (AUs) which are atomic facial actions. The FACs can decompose those facial muscular motions and identify the facial expressions in the Action Units. Following this, the AUs are then applied for any important decision-making procedure including numerous affective state recognition, basic emotion recognition among other complex psychological states.
The system manually codes the facial expressions through a set of specific guidelines. Currently, the FACS and Ekman’s basic facial emotion model are the most used methods in the field of vision based facial emotion recognition. However, it is important to note that the input for the FACS is static images of facial expressions which can lead to a lengthy and time-consuming process.

The pioneer work developed my Ekman has inspired numerous researches to conduct facial expression analysis through the use of videos and images. The facial movements and facial feature tracking were used to categorize the facial expressions. Some of the surveys conducted have provided an in-depth review of most of the research completed in the field of automatic facial expression recognition. Generally, four main steps are used to recognize emotions from facial expressions. These are the following:

1) Face extraction
2) Face detection
3) Feature extraction
4) Classification.

A very limited number of systems that recognize facial emotions can deal with both image sequences and static images. Pantic and Rothkrantz published a survey which emphasized how facial expressions can be automatically analyzed. The detection of the face occurred using a watershed segmentation which had markers that were extracted on the HSV (Hue Saturation Value) color model algorithm. A point base face model is utilized, and the characteristics are defined as certain geometric relationships that occur between the facial points in a defined small area relative to those facial points. A wide range of feature detectors is used for each of the facial features locations and the model feature extraction. Through the utilization of the rule-based classification method, the reported recognition rate is found to be 86%. In [26], a set of multi orientated and multi-scaled Gabor filters were utilized to
transform an image. Following this, a registered grid is obtained using the elastic graph matching method. This registered grid is sampled and combined into one vector as features. A 75% accuracy was achieved when this was tested on a database of approximately 193 images. An average recognition rate of 60% was obtained through the detection of the positions of the eye and lip using the low pass filtering and edge detection method in [27]. Wang and Guan (2008) have proposed a bimodal system of facial emotion recognition. An HSV color model is used to detect the face, and Gabor wavelet features represented the facial expressions. Through the usage of multi-classifiers, an overall recognition rate of 82.84% was achieved.

Despite that numerous methods are used to recognize facial emotions; the most used ones are face tracking, face detection, feature extraction, and classification.

2.2. Facial Expressions

The complex structure of the human face requires a great deal of attention. The structure of the muscles and their cooperation with other parts result in the face being one of the most emotional parts of the human body. It is challenging to predict the inner state of someone’s mind without being able to see their face. Thus, facial expressions play a dominant role in nonverbal communication.

The Universality Hypothesis asserts that the perception and the emitting of facial expression and emotions are identical regardless of the individual’s ethnic or cultural background. Charles Darwin initiated the original work that studied the facial expression and their consequences. He claimed that facial expressions could not be learned and pose an evolutionary meaning for the survival of people and therefore he described them as being innate. An observational study conducted by Paul Ekman was developed to establish whether humans around the world present similar appearances of their emotion. His research concluded that there were a certain similarity and a level of universality. Furthermore, Ekman conducted a study in New
Guinea on an isolated tribe and discovered the same signs of facial expressions that are used by other modern civilizations around the world. Also, Ekman presented them with “our” universal expressions, and the people of the isolated tribe were also able to recognize them.

Psychology differentiates between emotions and feelings. While the duration of emotions is limited up to five seconds, the duration of feelings can last for numerous hours. The emotions are even more distinguished as spontaneous and voluntary expressions. Voluntary expressions are controlled by the individuals and can easily be faked whereas spontaneous expressions are very brief.

2.2.1. Physiology of Facial Expressions

From a physiological perspective, facial expressions are a consequence of facial muscle activity whereby the muscles are also labeled as mimetic muscles or more widely known as the muscles of facial expressions. These muscles form part of the group of the head muscles which also contain scalp muscles and muscles of mastication which are responsible for the movement of the tongue and jaw. The facial nerve innervates the facial muscles and branches out in the face. The activation of facial nerves causes contractions which lead to numerous observable motions. Some of the most visible muscle movements are the block of movement of the skin such as the cheeks, lips, and eyebrows which can be found for example of the nose or on the forehead.

The human face comprises approximately 20 flat skeletal muscles as indicated in Figure 5. The muscles are under the skin, and they are attached to the bone of the skills and inserted into the skin of the face but not inserted into the bones and joints as is the case of other muscles which are responsible for other body movements. The muscles are close to the facial orifices such as the eyes, nose, and mouth. Dissimilar to other muscles located on the face, they are not able to move with the bones and
joints and often only move with the skin. Accordingly, some facial deformations can be causes which can lead to a variable facial expression which can represent emotions. It is important to note that the movement of muscles occurs in groups rather than independently and they take control of the orifices. Based on their location, the taxonomy is categorized into three main groups: orbital, nasal and oral.

The shape of the oral orifice can be altered by the oral muscles, which are responsible for all of the complicated mouth movements that allow the sophisticated form of the mouth, such as the depressing of the right and left corner movement of the cheeks, the encircling of the mouth, the elevation or the depression of the upper and lower lip and the angle which controls the mouth.

Figure 4: Muscles of facial expressions, taken from [33].
The nasal group controls the opening and compression of the nostrils. One of the nasal muscles is located between both of the eyebrows and poses as one of the most critical facial expressions as it can pull the eyebrows down and can lead to wrinkles appearing on the face of the nose.

The orbital group consists of three main muscles which are responsible for the movement of the eyelid and the protection of the eyes. All of these muscles are inserting the skin surrounding the eyebrows and form vertical lines of wrinkles between each eyebrow and the other.

2.2.2. Facial Action Coding System

Ekman et al. developed the Facial Action Coding System (FACS) which is a tool that is used to describe each facial muscle activity movement by a set of Action Units. A specific AU represents a particular component of facial muscle movements as can be seen in Table 2. Every motion in the face is described by a set of Action Units as can be seen in Table 3.

Table 2: Description of several AUs together with the muscle name [36].

<table>
<thead>
<tr>
<th>AU</th>
<th>Description</th>
<th>Facial muscle</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Inner Brow Raiser</td>
<td>Frontalis</td>
</tr>
<tr>
<td>2</td>
<td>Outer Brow Raiser</td>
<td>Frontalis</td>
</tr>
<tr>
<td>6</td>
<td>Cheek Raiser</td>
<td>Orbicularis oculi</td>
</tr>
<tr>
<td>7</td>
<td>Lid Tightener</td>
<td>Orbicularis oculi</td>
</tr>
<tr>
<td>9</td>
<td>Nose Wrinkler</td>
<td>Levator labii superioris</td>
</tr>
<tr>
<td>10</td>
<td>Upper Lip Raiser</td>
<td>Levator labii superioris</td>
</tr>
<tr>
<td>11</td>
<td>Nasolabial Deepener</td>
<td>Zygomaticus minor</td>
</tr>
<tr>
<td>12</td>
<td>Lip Corner Puller</td>
<td>Zygomaticus major</td>
</tr>
<tr>
<td>13</td>
<td>Cheek Puffer</td>
<td>Levator anguli oris</td>
</tr>
<tr>
<td>14</td>
<td>Dimpler</td>
<td>Buccinator</td>
</tr>
<tr>
<td>15</td>
<td>Lip Corner Depressor</td>
<td>Depressor anguli oris</td>
</tr>
<tr>
<td>16</td>
<td>Lower Lip Depressor</td>
<td>Depressor labii inferioris</td>
</tr>
<tr>
<td>17</td>
<td>Chin Raiser</td>
<td>Mentalis</td>
</tr>
</tbody>
</table>
2.3. Related Works

The analysis of facial movements is growing in popularity in the field of research over the past few years. Some of these applications used to conduct facial study are expression analysis, recognition, and the simulation of the animation industry. Humans are naturally able to recognize the emotions which correlate with certain facial expressions. Ekman and his team of colleagues have been able to cover all of the six basic expressions that are identified using the FACS in fixed images.

Some approaches have been developed using videos and still images for facial recognition systems. Black and Yacoob are the first to introduce local paramedic motion representation. Chang et al. created the low dimensional manifold technique for the representation, training, and identification of facial expressions. They used the Lipschitz embed system alignment for the facial feature in the construction of the manifold. Some of the rarer forms of facial expressions were studies by Kumano and his colleagues using the templates of variable intensity, and they were able to state differences between numerous classification templates. Following this, the usage of the Support Vector machines and Adaboost classifier rarely gained benefits.
Sometime after this, Torre’s methods were recognized for their graphical spectral techniques that were somewhat unique in this department.

The expressions of the human face are the visible activation of the human’s state, their cognitive activity, and their personality. Along with other gestures, facial expressions can provide the listener with an excellent understanding of the intended meaning behind the speakers’ speech which can lead to effective communication.

The FACS that was proposed by Ekman can describe the distinct facial movements. Through the parameters of action that is designed to classify the human emotions is what the FACS is based on. It was found by research conducted by Mehrabian (1971) that facial expressions provide a 55% effect of a given message, while the vocal part contributes to 38% and the verbal only has a 7% contribution. Numerous applications of facial expressions are used including image compression, medicine and psychology.

Due to facial expressions playing a significant role in presenting information about an individual, they can play an even more significant role in the interaction between humans and computers. The automatic emotion recognition derived from facial expressions may act as one of the components of the natural human-machine interface. This type of interface provides services which require an in-depth understanding of the emotional state of the users. For example, using the interfaced in some robots to be able to recognize expressions made by humans can pose as a significant contribution to the creation of more intelligent robots.

Since 1990, numerous works have conducted with automatic emotion recognition from images and videos. [54] OF (Optical Flow) was used by the researcher to provide an estimation of the facial muscle movements to recognize the expressions of the face. For the first time in this field image processing techniques were used to recognize facial expressions.
A flexible appearance model was used to locate facial features, recover pose, coding, and reconstruction and to identify individuals with a particular image. A collection of OF models was used to track and recover both rigid and non-rigid facial movements. Certain simple head movements and eye blinking were also recognized by this system. The researchers used similar rules as in [56] to classify the emotions and computed the OF. The OF areas of the face were computed, and a radial based function network was utilized on the facial emotion detection software. The OF flow processing was used to calculate and classify facial emotions. An OF algorithm was initially used to provide a calculation of the velocity vector. Following this, a 2D Fourier transform was used for the vector in the mouth and eye area. Finally, the HMMs were used to recognize facial expression from an image sequence for some individuals. It was concluded in these experiments that the proposed systems were able to recognize the six basic emotions in nearly real time.

Even though there are numerous achievements for recognizing emotions derived from facial expressions, there are still some difficulties faced in addition to limitations as a result of the complicatedness of emotional expressions, particularly during conversations. Various benefits result from the recognition of emotions that are derived from facial expressions. These are the following:

- It is one of the most natural methods used to observe the emotional state of individuals.
- There is a wide range of databases that can be used to conduct research.

Some of the disadvantages include:

- The information found on databases can be misleading at times,
- The recognition results are heavily dependent on the image or video quality.
In recent times, there have been considerable efforts conducted on the implementation of emotion recognition systems using accosting information and facial expressions. [60] has proposed an emotion recognition system that combines both video and audio data using the rule-based system. The authors have described the use of statistical techniques and hidden Markov Models in recognition of emotions. They used the prosodic characteristics of audio and the maximum distances and velocities of videos between the five different facial points. When they used both of these collectively, they witnessed a higher performance of emotion recognition.

2.4. Factors of Facial Emotion Recognition

Vision-based emotion recognition is mostly concentrated on expressions of the face because the face plays a vital role in expressing feelings and perceptions. There is a wide range of approaches that have been applied in this field. Faces are considered to be nonrigid, and they have different colors. Certain facial facials are quite rare and may be unsuitable for the pattern recognition. The lighting and background play a significant role in determining the recognition rate of facial expressions which makes the identification of emotions derived from facial expressions more complex.

Currently, some facial expression recognition approaches have been formed on 2D data such as static images and 2D video sequences. This research will also utilize the two-dimensional approach to conduct our emotion recognition from a facial expression. The recognition of facial expressions requires four key steps which are: face detection, face tracking, feature extraction and the classification which are presented in figure 6. On the other hand, speech emotion recognition requires three main steps which we will discuss in the next chapter. The system operates on image sequences that have been taken from a video camera. In training and testing processes, static images are used; however, it is important to note that the interaction with a system is designed for video analysis.
2.4.1. **Face Detection**

Although there are numerous forms of input to the emotion recognition system, images that contain faces remain essential for intelligent vision-based human computer interaction. Face detection strives to identify all the image areas which include a face regardless of their three-dimensional position, lighting condition and orientation of the particular image. The issue that arises here is that faces are non-rigid and possess a high degree of variability regarding their shapes, sizes, colors, and textures. There have been a large number of face detection techniques created. The authors have categorized the detection methods into four main categories: Appearance-based methods, Template matching methods, feature invariant approaches, and knowledge-based methods.

Many researchers commonly use Feature-based method, it is important to note that their disadvantage is that they are highly dependent on the noise and the illumination of the image. As for appearance-based methods, they are learned from a set of training images. They predominantly rely on the techniques used in statistical analysis to identify the applicable features of facial and non-facial images. Many of these distribution models are used to model the learnt features of face detection including Eigenface [70], Gaussian distribution [71], Neural Network [72], SVM [73], Naïve Bayes Classifier [74], HMM [75], and Information-Theoretical Approach [76, 77]. Despite their computational load, the appearance-based methods are superior to the other methods.
Knowledge-based facial detection methods have also been known as the rule-based methods. The extraction of the facial features occurs from the input image and following this; the faces are then identified based on these coded rules. The issue with this method is that is can be challenging to translate the human knowledge into well-defined rules. Moreover, it can also be difficult to extend the approach so that it can detect faces for various poses.

In comparison to knowledge-based methods, the feature-based method utilizes the invariant facial features to detect a face. Some of the features include the hairlines, mouth, nose, and eyes. Following this, a statistical model is then built to verify the existence of a particular face.

In [79], The authors have proposed a frontal face detection system that utilizes greyscale images through the Hrr Feature-based Cascade Classifiers. It can be used in real time to detect faces, and it is available for all researchers on the OpenCV tool. The works of the authors have been extended so that they can handle multi-pose faces using cues from skin colors to result in a decrease in the computation time. Currently, the Viola and Jones facial detection method is popularly used by researchers and has also been used in our research.

2.4.2. Face Tracking

Face tracking is used in the video frames to pass over the detected face backward and forward of each specific frame. Generally, two methods have been used in face tracking; they are using the face detectors as a tracker running indefinitely and the second approach is the development of a face tracker apart from the face detector. Some of the most widely used tracker models are AAM and ASM.

In the face tracking, the shape of a face must be reconstructed to fit the target face image. With regards to the ASM based method, training images manually labeled are mostly used. This approach starts by searching for the salient points to fit the model
in the image, and these points are updated at each frame. This method is also called Smart Snake. As for the AAM method, it also used a training phase, but it utilized both the appearance and shape information of the target image. Accordingly, the ASM method is faster because the AAM method utilizes all of the image’s data. The original algorithms are used for grey scale images, but they can both be extended to cover color images as well.

2.4.3. Feature Extraction

Extracting facial features plays a significant role in recognition of emotions that are derived from facial expressions. The lips, nose, and eyes are the most significant facial features. Some approaches have been used for the extraction of facial points from videos and images of faces. The extraction of features is closely related to the detection of the face since, in the detection of the face, there are some of the features that have already been through the extraction process. In that situation, we just introduce the different concepts of feature extractions.

There are generally four types of methods used for facial feature extraction. These are:

- Color segmentation-based feature extraction
- Template-based feature extraction
- Geometry-based feature extraction
- Appearance-based feature extraction.

Geometry based approaches extract the facial feature through geometric data such as the size of facial components such as the mouth and eyes, in addition to their relative position of the face. The template-based feature extraction matches the facial feature to a previous template design through the use of suitable energy functions. The color segmentation method uses the skin color to divide the face, and the non-skin color area will be seen as a selection for the mouth and eyes. The appearance-based
approach provides the necessary components of the face in addition to the texture of the face such as wrinkles and spots.

2.4.4 Emotion Classification

There are numerous research surveys on the classifiers of facial emotion recognition in the literature. In the case of static images, facial expressions are classified based on the tracking results derived from each of the images. There are two types of classifiers which are generally used to classify facial emotions which are Naïve Bayes classifiers and Bayesian network classifiers. The Hidden Markov Model-based classifiers have been suitable for the recognition of facial expressions and have been commonly utilized in this field. A multilevel HMM has been proposed for the automatic recognition of facial expressions derived from video sequences. The type of classifier that is used in recognition of emotions’ step is dependent on the facial detection method and the extracted features.

2.5 The Case for Affective Research Using Technology

The 19th Century witnessed studies on the recognition of the effects of emotions through facial expressions focus primarily on distinguishing the simple variances between positive and negative states which is substantially due to the challenges of the time, such as technological processing power. At this time, there prevailed a theory which stated that the muscles contracted differently as the subject experienced positive and negative emotions, a two-dimensional approach. More systematic and complexed observations have occurred in the last three decades thanks to advances made in scientific fields like psychology, neuroscience and cognitive science. These studies have married the use of computers, integrating them as a powerful classifier of these different states. The hope is that by doing so, further illumination may be created upon a new field of the integration of humans and computers.
Furthermore, there have been notable advances in the measurement in the singularities observed in external emotional signals – an example being Ekman’s Facial Action Coding System. It monitors select facial movements and reconciles them with the emotion. There have been some breakthrough studies in this regard by the likes of Davidson, Stemmler, Harrigan and Fontaine, each playing a pivotal role in the movement towards better understanding in the field. However, the studies in question present the notion that emotions are intrinsically complexed and often resist systematics. It shows itself through patterns. The last two decades have seen a shift towards getting computers to identify emotional displays by its facial affect. The implications of this feat reach far beyond academia.

2.5.1. The Current State of Technology: Machines with Emotions and Personalities

There has been a marriage of computer science and psychology to create systems that allow for the identification and prediction of emotions. Picard designed sensors that, when applied to the skin, identified psychological changes. Furthermore, they analyzed the ‘patient’s’ psychological condition and gave biological information based on the reading. For example, to improve an individual’s mood, it would prescribe specific practical measures. There has been more to explore down this style of research, for Paul Ekman also sought to correlate expressions and emotions. He labeled these expressions as social signals that aid in communication between persons. Together with Friesen, there came about a more precision based causal relationship between specific emotions and their emotional counterparts.

Technology may be used to analyze thoughts and intentions. Before this can occur, the computer must first go about recognizing the presence of human faces in any given image. After this is verified, the processor may then search for any signals which correlate to a specific emotive effect, as it was programmed to do. Further classification may be available depending on the intention of the research. However, this is easier said than done, as the link between computers and emotional recognition
has dominated the literature for the last 30 years. Some argue that there may be too many variables in any given image to formulate a well-conceived result. Such difficulties may lead to obstacles that prevent the combination of technology and psychology, as human emotions are subject to so many differences at any given point. The result may lead to a calculation that is too difficult to define at this point. Variables include variations in location, scale, pose, expressions, orientation, lighting conditions etc.

2.5.2. Facial Detection Technology

Initially, the presence of a human face must be recognized in the image by the technology, a technique that relies on a wide variety of aspects from the relating literature. One such aspect is a knowledge-based technique which relies on a checklist of regulations to recognize the individual face. These rules act to define a face, thus identifying to whom it belongs. Another is described as a feature-based technique that identifies the face by recognizing distinctive critical features. Lastly, in the research one would find mention of a template-based approach, a method which identifies based on pre-loaded patterns, textures, and structures. It is worth noting that appearance versions of facial recognition are widely regarded as superior, even though they may require additional RAM.

There are, as stated before, a wide variety of variables that may affect the reliability and accuracy of the technology, variables such as lighting, color differences, size and textures, and non-rigid structure. Position, existence or absence of objects all play an integral role in the recognition process.

Were an image to be scanned through a window, with various scans taken at a differing angle, by facial recognition technology, there would be dramatically different results that serve to prove the point that facial recognition is showing to be a costly research topic. Various problems have not ceased to exist, some of which
may be solved with a training mechanism of the sort. The demand for such a classifier is monumental, as these devices (classifiers that produce the least false recognition) are advantageous.

A well-designed system may aid in solving the budget problems involved with facial recognition development. There are even algorithms that adapt the technology, so that face detection is made more comfortable with the progression of time. Each algorithm may be assigned to a general task, for example, identifying the user of a phone. Viola and Jones are one such provider of a learning algorithm, an anterior facial classifier that is most applicable for everyday purposes. It is known as a VJ algorithm, and it is available for the use of furthering the research of live face detection. Further applications of algorithms of similar stature are possible from the Machine Perception Toolbox.

However, the software may come with negative implications, for at certain times the rate of false detections may increase. As an example, if 507 faces were taken as a sample, it would result in 150 false positive identifications. The classifier has fared well in studies with the use of the algorithm, but it is still not at the point where one could consider it fit for useful applicability. The literature, therefore, points towards more classification filters that hone down the skill of the classifier. An example may be skin color criteria.

### 2.6. Studies in Facial Expressions

Pre-existing scientific theories have led to the advent of intelligent computer systems that can copy human emotions. Some might even attribute the intrinsically organic concept of learning to these machines. Automated emotion recognition, in particular, has made dramatic advances, especially regarding the identification of emotional facial expressions. Moreover, the field remains the point of research for many, with many stores of data collected such as the Illumination and Expression Database, the
University of Oulu Physics-Based Face Database, Karolinska Directed Emotional Faces, the Japanese Female Facial Expression Database, the CMU Pose, the Extended Cohn-Kanade Database, the Yale Face Database, and their many derivatives. The information contained within these databases show some persons in various stages of emotional displays. These emotional displays are captured in slightly different manners, from an environment containing semi-natural stimuli.

2.6.1. KDEF Database

In 1988, a database was created for the non-commercial storage of facial expressions, namely the Karolinska Directed Emotional Faced (KDEF). The KDEF is now seen as the standard for such operations, as it was a thorough study. It was made in 1998 by Anders Flykt, Daniel Lundqvist, and Professor Anne Ohman at Karolinska Institutet, an organization based in Stockholm, Sweden. It is a database which contains 49000 images, taken by specialized equipment, so that a variety of facial expressions may be captured. The snapshots are made from different angles and provide an example of a range of different emotional facial expressions, as per the example below:

![Figure 6: Male subject is showing all the seven basics from the KDFE dataset.](image)

2.6.2. JAFFE Database

The JAFFE Database, with the research of Michael J Lyons, sought a method that gave expressions a quantitative 5-level scale as the emotions captured were never outrightly classified as just one pure emotion, but rather a complexed pooling of many at any particular time. JAFFE makes use of a total of 219 images per subject
captured within a controlled environment. The sample size consisted of 60 Japanese females who were subject to a psychological experiment so that they would display the needed differing emotional expressions. Expressions were rated as per the dominant emotion displayed by the 5-level scale. However, the image resolution is comparatively low when compared with similar research. Further, to its discredit, this database contains a diminutive sample size of just ten people. Below is an excerpt from the JAFFE Database.

![Figure 7: Examples of images from the Japanese Female Facial Expression database.](image)

### 2.6.3. Cohn-Kanade Database

97 subjects were chosen from the university for the Cohn-Kanade Database. Their ages sat between the spectrum of 19 and 30 years. Within the group, a vast number of non-white ethnic groups were represented (African American, Latin American, etc.). In aid of the study, various recording tools were given to the subjects such as high-quality cameras and video recorders. These were used to record participants at multiple angles – a front pose and a pose with the alteration of a few degrees. Neutral expressions were recorded as an initial sample from which to draw a comparison with any later facial changes induced by emotion. Furthermore, for the sake of capturing perfect results, participants were shown effective expressions before the study. This ‘coaching’ made use of Action Units in full view of a live, frontal camera.
The images captured were digitized. The technology in use was of sufficient standards to ensure a study of accuracy (8-bit grayscale values and 490-pixel arrays). Moreover, these images were formatted to png and jpg, and are labeled according to the corresponding VITC. To further the cause of determining the exact consequences of the correlation between emotions and facial expressions, the researchers employed FACS, Facial Action Coding System. This technology helped explain the expressions of the candidate. Here is an example of the Cohn-Kanade Database:

![Figure 8: An image sequence of a subject expressing 'Surprise' from the Cohn-Kanade Facial Expression Database.](image)

### 2.7. The Current Literature on Affective Dimensions & Cognition

The previous 50 years has captured a fair amount of research into the questions concerning the effective power emotions have on human expression, and many theories have been proposed along these lines that began in the 1960s with Magda Arnold’s assertion that explained emotional expression as a derivative of external stimuli provided by a state of being. This theory hoped to demonstrate the elicitation of emotions. To this effect Frijda, too, that the kernel of emotion is obtained by the sensory acknowledgment of the presence of either pleasure or pain in any situation. However, Ekman further argues that the facial expressions derived from emotions are a social signal to better communication between persons. A collaboration of Ekman and Friesen would be the introduction of a more precise expounding of the theory, an explanation that takes into account the evolution of intensity of expressions. Also, there was an observation of the nuances involved in the emotional
expressions of humans. Subject to geometric analysis, one has to conclude that the results have been a description that is mainly hypothetical, but it does sit within a continuum. The literature surrounding these studies will always be tricky as the object of study will always remain an endeavor that is subjective.

Emotions have, for some time now, been given the responsibility for certain physical occurrences within the physical human body. They correlate with high blood pressure, adrenaline and heart rate, amongst others. Anger has proven that it does provide for key survival mechanisms in both humans and animals, as it is a reflex that seeks to dispel any threat or danger to one’s being. Therefore, it is natural that it affects one’s physiological state. However, as stated before, all emotions that are experienced are done so in a subjective manner and a unique environment. Each experience is not always comparable with another. All of the information above is mostly based on Arnold’s “appraisal theory” from the 1960s.

2.8. Facial Action Coding System

Databases, as discussed earlier in this literary work, have been created to further the research of facial expression recognition. However, there have been other methods employed that achieve a similar end. There is a hope that these methods would provide greater clarity within this particular stratum of study. An example of this would be the Facial Action Coding System, developed by Ekman and Friesen. It can recognize and quantify facial muscle changes in an objective manner. In 1965, it was revealed that some expressions derived from emotional sources that are universal and cross-cultural barriers, such as anger, fear, disgust, sadness, and enjoyment. To develop the technology, the team studied how emotional fluctuations played a role in the muscular changes of the face. Ekman et al. classified these muscle groups as Action Units. By doing so, FACS is easily able to use economical portions of facial muscles to provide an accurate reading.
Summing all the AUs would give a figure of 11. There are similar movements amongst 30 of these of these AUs, with 2/5 from the upper face and 3/5 from the lower face. These upper and lower facial AUs are subject to connection via connective, muscular arrangements. Below is a demonstration of the upper face and lower face changes and combinations. One can see how the facial of the AU contract and released by the corresponding emotion. Ekman et al. used this AU system as the basis for the Facial Action Coding System, a system which considers emotions as separate and a state that is discrete.
Figure 10: Upper face Action Units and some combinations [106].

Figure 11: Some lower face Action Units and some combinations [106].
Facial Recognition and Computer

FACS is a high maintenance system as the coding of Action Units, as minimal functionality can only occur with 100 hours of assimilation in human interpretation. Plus, the assessment of the data obtained within each minute of recording requires an hour for sufficient scoring to be acquired. As a result, the manual effort required in the FACS process renders certain limitations upon this method of research. There is a continued search for a more automated computerized facial recognition system. A possible effort in this regard has led to the research of Mase, an early version of an automated system. More recently, the introduction of feature extraction from images which makes use of pattern recognition models in similar images has been used. It can identify general patterns but not identical images.

Image Pre-Processing Methods

The latest versions of facial recognition have become so advanced that they are looking to be applied to the personal environment. The adaptation into the everyday life of members of the public is not far from reality. It is now far easier to achieve signal processing, data acquisition, and analysis systems. What was once in the in the form of signal transduction in which not only the input of an image but the eventual output may be some altered version of the original or some information thereof. A hefty filtration is made necessary by facial classification software, as the area that it scans is proportionally small. A higher level of precision is required compared to facial detection technology, meaning the increased need for RAM and computing power. The early editions of this technology were able to recognize faces on still images alone, as it evaluated the RGB value of every pixel to derive a solution from it. Still, this required many computations. It would be inefficient to apply such a system in the future.
Furthermore, as we now have access to greater computing speeds in a small and compact device, there should be significant scope for applying this technology in more practical measures. Therefore, we currently have the Haar-filter system, a technique that makes use of quadrilateral focus on an accurate location in a recognition area. In turn, this allows for pixels to be examined and the differences to be classified into subsets. For example, eyes would be classified into one subgroup, so too with noses, cheeks and mouths. These regions would have virtual borders placed around them, and the filter would then scan the surface area within the bounds of these digital borders. Within this surface area, the filter then identifies unique features and personal marks that act as an identifying mechanism.

2.9. Facial Expression Technology and Smart Phones

Smartphones have recently caught the eye of the market due to their facial recognition technologies, with certain smartphone firms testing the technology as a feasible idea for the future. Windows Mobile, for example, has run multiple tests in the facial recognition software domain. Moreover, they have done so in a real-time manner. Facial expression recognition has become increasingly popular as it becomes more and more in the interest of the general public. The net result has been an increase in momentum of studies in the field, an area that is now subject to special advancement. Machines becoming aware of human emotion will cause the quality and success of this advancement to accelerate because robots will be able to progress from a lesson in synchrony with human readiness.

Historically the students of facial expression have not been the interest of computers and IT. That title has belonged to medical practitioners, artists and actors. However, the latter half of the last century has seen dramatic advances in the innovations of computer vision and versions of AI that has allowed those within the computer studies environment to contribute to such an endeavor, starting in the early 1990s and continuing to the present day. It has evolved into a popular research topic. More
recent in roads have to lead to higher human interaction classification – more than just the facial and emotive recognition. For example, this technology may be used to analyze autistic people who struggle to show emotion fully. The technology should be able to help integrate such individuals into a group environment.

The amounts of uses for the technology that classifies emotional expression seems almost infinite at this point. Today, one may even enjoy these types of technologies on a mobile device, a possible nod to the fact that people see an immense allure in constructing virtual worlds. There are some applications available on websites and app stores that help explain emotions. These may be used for emotional recognition exercises. However, most of these applications require the user to identify themselves according to a particular, prescribed template that does not allow for accurate emotional recognition, as the emotive state is exceptionally complexed. Therefore, a proper classifier should be reactive rather than have a set standard that one must adhere to.

The Reason for Facial Recognition Adaption to Mobile Phones

The market is calling for the use of facial recognition technology to be adapted to mobile phones and PDA devices. Therefore, as there is scope for commercialization, much of the research has been geared toward this goal. Underneath is a list of reasons for the adaption of the technology to mobile devices:

- It acts as a protective mechanism, as facial recognition is by far superior to a passcode.
- Facial recognition software has entertainment value
- It allows for image correcting, such as blurs or distortions
- The software may be used to track criminals by matching the policing databases with the facial recognition software. This may be particularly relevant in the case of mobile phones, as it has become a necessity of modern culture.
In mobile phones, a system called the biometric technique is used, a method that is used as part of the facial recognition system of other methods. In the biometric system, statistics are used as a collection for the classification of a wide variety of items – from actions to geomorphology. It is a system that has its benefits, as secret codes or signatures are now redundant prerequisites of access to devices (or anything else, for that matter). If the system is robust and flexible (as any suitable system should be), only the desired user may have access to the contents of the mobile device.

The Best Facial Recognition Apps on the Market

As it was stated in previous segments of this paper, there are a wide variety of facial recognition applications available for smartphones. These are primarily for the sake of recreational values, but other genres in which the technology is applied includes artistry. MARA3D Facial Expressions allows for access to useful and fully customizable facial expressions. Further, worthy mentions are given to Body Language Expressions, Fun Face Changer Extreme free and Vola Friends Application for Children. Snapchat makes use of a ‘face swap’ system.

The applications that make use of the services of facial recognition tech have evolved dramatically over the last few years. However a gap between the app that is required and the ones that are currently available remains. This distance is within distinct areas, such as the confusion created when matching body parts and attempting to apply it to the application. Further technical faults lie in the amounts of languages that such apps contain, as the perfect app would include all available language options. The studies show that the software relies much too heavily on self-reporting, as it uses physiological data from a particular sensor alongside self-reporting to provide a general diagnosis of a very specific emotion and it may not prove to be the most accurate of methods which carries a especial concern to the psychiatric field.
People who have autism may not be able to showcase general emotions and would, therefore, struggle to use such technology. The direction of research must take into account full expressional nuances to provide the results desired. With the pace at which the research is gathering and the direction of the community, this may still be possible.

Tips for Writing Smartphone Programmed

The creation of an application is a complexed endeavor, one which requires a specific set of skills to produce a successful product. Experience in programming, coding and a clear vision of what is needed is essential for success. Many applications are commercialized in a variety of sectors. However, to ensure success, a recommendation of developing within a sector that is known to the creator is to be adhered to. Do not write an application in gaming if your knowledge lies in spelling systems and autocorrect software. Furthermore, the first few months after development require tremendous stores of capital, as the general trend shows that an application does not generate sufficient income for the first few months after release.

There are many ways to go about such an endeavor, but for the sake of a comparative tool, below is a representation of making a workable programmed. From this, one may derive an educated opinion as to what and how to develop an application.
The market leader for the application development market is undisputedly Java. It is an independent platform, one that does not contain licensing. Furthermore, it provides an integrated development platform for users coupled with data encapsulation and the utmost quality of security. The data hiding is most secure in Java compared to other languages. To add to its credit, Java has become the language of choice for Android. Robot enables the prerequisite tools and APIs, which in turn allows for the Robot OS to use the Drink Programming language. Remarkably, Drink is listed as the champion of the TIOBE Programming Accord Forefinger. There is a reliance that programmers have on these languages that is hard to underestimate, as it allows for new Beverage applications.
CHAPTER 3: SPEECH EMOTION RECOGNITION

3.1. Introduction

The role of speech is to communicate information in a manner that conveys information to another person. Within this is emotional information – communicated by both the diction that comprises the speech and the manner in which that diction is spoken. There is content that delivers the same objective message which may be stated in a variety of ways, using techniques that convey entirely different connotations. There is a broad acknowledgment that there are two types of communications within speech [115]. These may be classified as linguistic and paralinguistic. On the one hand, messages are given through the medium of language. Someone may say the actual words, “I hate you.” Likewise, there are paralinguistic accompaniments to the linguistic elements used. These are implicit and must be deduced by the recipient of information. For example, the person who states the words, “I hate you,” is likely to have a frown on their face and use bodily gestures in such a manner that it is evident that they mean what was spoken. Such techniques seem to reinforce the emotive impact of utterances.

The explicit messages are comprised of the objective contents of the speech, the concept looking to be conveyed. An implicit message is a manner in which those concepts are portrayed. It may include emotional undertones, grammatical structure, emotive language, etc. As stated in the previous paragraph, these may act as accompaniments to the verbal diction itself. However, there is recognition that it may be contrary to the linguistic message. Failure to comprehend the emotive values in an utterance inevitably leads to misunderstanding. The concept, itself, may not even be understood as it was intended to be. It remains difficult to develop concrete rules [116] surrounding the linguistic content and emotion contained within speech, as they remain reliant on language. Therefore, these studies on the base model for the
observation and recognition of speech emotional recognition further show the reliance of emotive speech on language.

An emotional state in speech is largely based on phonetic elements and acoustic features. These elements comprise what is termed as the implicit message. An example of this can be seen in one of the main determinants of the index of arousal: the pitch function in the human voice [115]. Today, there has been a dramatic shift towards the development of literature, through empirical studies, so that further illumination may be given into the field of implicit messages within speech. There is more yet to be uncovered, such as the unknown factors in the range of acoustic features observed that are derived from the same emotional state. With a better understanding of emotional recognition from audio information comes a list of advantageous implications: lower costs and the options of a large sample size. Disadvantages include the dependence on selected languages and the lack of accurate identification that comes through the observation of facial expressions.

The field of emotional identification that is contained within speech has formed an integral of research, with multiple means to identify the emotive constructs included within speech. This is done through critical analysis of the acoustic and linguistic data that is found in any given form of speech. There have been successful multiples of different kinds of emotion recognition – two being an online approach and an offline to contrast. The former has rarely achieved successful outcomes.

3.2. Basics of Emotions in Speech

With speech, there is a certain amount of emotions that remain unspoken as they are almost always automatically recognized. The human brain naturally and effortlessly recognizes some of these emotions in a manner so elementary that one would think past classifying them. However, for precision, these and all emotions need to be identified to develop theories and furthering research. Psychology is the primary
science which hosts the methods upon which the information about emotion recognition has been derived. Psychology includes armies of complexed models in the effort of this.

Next, one must endeavor to answer questions about the origins of these emotions, the place where they may be observed. The answer is that these expressions mainly lie in language and facial or body gestures. It is worth noting that while the content of dialogue may contain information that leads to confusion about emotional states, the following is never subject to human interference and often proves a better indicator of emotive reactions. Consider a scenario where someone may state: “I am fine.” If that individual were to say it with a trembling voice and a worried expression, one would know immediately that the opposite is true. The language contains the likes of acoustic and phonetic information. Technology has evolved to the point where we may now deduce further details from bodily signs such as heart rate, adrenaline levels or perspiration.

Concerning these concepts, a more detailed look at the automatic emotional identification contained within speech is still warranted. To do so, a broad design system must be presented. This extends from audio segmentation to classification. Furthermore, this may require much-needed features like correlation analysis and the extraction of relevant characteristics to construct an accurate thesis.

### 3.3. The Expression of Emotion

Now that these emotive as aspects have been adequately described, further thought must be given as to where these emotions may be located, and therefore, observed. Because emotions are bodily reflexes, they play a part in all physiological operations – language and speech for example. These two provide a focal point for this discussion, compounded with a mention of modalities that are outside the realm of linguistics, like facial expressions and ‘body language.’
3.3.1. Language and Speech

Within every facet of language emotional data is found. Such emotive information is communicated by what is said and the manner in which that information is presented. Sometimes the way in which the information is given is more important to understand than the message itself, as per the previous examples of an effective state contradicting the spoken dialogue. One must consider all aspects of language, from pragmatic to acoustic techniques, which are implemented. By doing so, one may observe the following: the intention of a speaker is closely correlated with that of the emotional state of the originator [118]. To this end, research has been undertaken proving that figurative devices (such as metaphors, irony, and sarcasm) can communicate emotive information that is not as easily defined as literal statements [119]. They often prove to be a subtler version of an outright statement.

An example of this is the comparison between the phrases “I have reached the boiling point” and “I am angry.” The former is by far the most precise. Literal statements form the most obvious displays of emotion, containing statements like “I am happy” as obvious emotional expressions [120]. Added to this, literal expressions may also be used as devices that display more intricate emotions. These statements can be used to say one thing, whereas the speaker means the opposite; it is a use of language that is easily manipulated. Fries [121] has comprised a list of speech whose syntactical construction have an underlying emotive meaning. As an example, one can observe that the phrase “This is very good” conveys much more emotion than the phrase “That is good.” Built on this is the notion that specific aspects of speech denote higher levels of emotion – prefixes like “super” for example. Techniques like these have a positively intensifying effect on the emotions perceived by the recipient of the message. However, this may be used in an ironical fashion, which again changes the emotions recognized.

The tone that accompanies a message is a further indication of emotion. A tone is created by a combination of phonetic and acoustic techniques that combine to
produce the desired pitch. Therefore, a change in tone is closely correlated to a change in emotion. The primary focus of this study shall remain on the recognition of emotion through an acoustic medium within the dialogue.

Psychological studies have proven the most important in this regard, with the results thereof showing that pitch, intensity, speaking rate and voice quality to be the most critical aspects of tone. Murray and Arnott [122] have endeavored to identify the most cryptic of acoustic correlations concerning its connection to emotion. Below, in Table 5, is a summary of these studies, an observation which shows prosody and the quality of voice to be the most influential factor in the human classification of emotion. The results draw particular attention to pitch and intensity, where a high emotional index is indicated by high pitches and high intensities and vice versa.

Table 5: Variations of acoustic variables observed in emotional expressions compared to neutral speech. Synopsis of Murray and Arnott [122]’s summary table.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Pitch</th>
<th>Intensity</th>
<th>Speaking rate</th>
<th>Voice quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>Higher mean</td>
<td>Wider range</td>
<td>Abrupt change</td>
<td>Breathy Chest tone</td>
</tr>
<tr>
<td>Joy</td>
<td>Higher mean</td>
<td>Wider range</td>
<td>Higher Slightly faster</td>
<td>Breathy Blaring</td>
</tr>
<tr>
<td>Sadness</td>
<td>Lower mean</td>
<td>lower Narrower range</td>
<td>Slower</td>
<td>Resonant</td>
</tr>
</tbody>
</table>
At first glance of Table 5, emotional recognition may seem uncomplicated. However, this table only takes into account the studies of one author. There is literature that does not entirely agree with the prescriptions of Murray and Arnott [122], Cowie et al. for instance [123]. The disagreements are likely due to the varying intensities of emotion, and further due to its variable nature. Emotions are highly subjective and the expressions of that vary. Therefore, there can be no direct mapping, as will be demonstrated later in this paper.

### 3.3.2. Extra-linguistic Modalities

To bring about studies of further accuracy, body language, and facial expressions must also be factored into the equation. The field of psychology has proven these techniques to be a far more reliable indicator of emotion than that of just linguistic nature, as the author of the words themselves has better control over the content of the message than the expressions accompanying it. Therefore, words may be used as a decoy, rather than the conveyor of emotion. From the list of visual indicators, emotion is most easily recognized through facial expression [124, 125] and also often through bodily gestures [126].

Furthermore, the physical state of any human being is directly correlated to that of the emotional state. These physiological aspects are quantifiable through a variety of measures: skin conductance, respiratory frequency (as demonstrated by Wagner et al. 127 and Honig et al. [128], pulse, and electromyogram. These provide the medium for obtaining information about emotions that are not subject to human manipulation. Another such mechanism would be the event-related brain potentials obtained from...
an electroencephalogram [129]. The technology, at this point, is an inconvenience and purely invasive. However, with the evolution of technology, this should see rapid improvements. Possible ideas involve including the technology within a shirt.

Surprisingly, emotional recognition may even be aided by the observation of the interaction between humans and technology. Schuller [130] observes that the patterns of mouse clicking may be used to provide insight into the emotional state of the user. Users' expressions may be enabled or expressed by these simple interactions with basic technology. For instance, Rehm et al. [131] have shown links with emotion and Nintendo’s Wiimote. Also, Sentoy (a doll that has sensors which allow emotional interaction with users) has been used to identify the presence of emotion. It was introduced by Paiva et al. [132] and Lv et al. [133].

There is a respective amount of data concerning emotional states that may be deduced from singular modalities. However, this significantly limits the accuracy. Therefore, as stated by Hudicka [134], the highest accuracy is derived from the observation of multiple modalities at the same time. A particular scenario determines the types of patterns that must be observed. Certain modalities easier recognize some emotions. Sometimes the circumstances do not allow for the observation of all modalities; for example, the dialogue which occurs over the telephone does not allow for facial and bodily gestures to be taken into account. Moreover, machines are not able to provide human-like emotional states, as they are only able to hold one emotion within its scope. Despite this, they play an essential role in fully understanding individual modalities.

### 3.4. Automatic Emotion Recognition from Speech

Emotions have played a long-standing role in the realm of psychology, as alluded to in the previous sections of this paper. In contrast, the field of informational technology dismisses this area of research as unimportant and even contradictory to so-called “proper” technical application. In the 1990s, Picard introduced the topic of
Affective Computing in a revolutionary work [135] with the same title. Her work proved the basis for the consideration of emotional studies using technology as the means to an intelligent interface, allowing for the identification and reacting to the emotional states of consumers of technology. The applications of this provide better efficiency and convenience. The last ten years have seen the rise of many solutions to affective computing. The list includes the following: Cognitive human variables exploitation, applications, databases, usage scenarios, and speech resources.

3.4.1. Overview of a Statistical Speech Emotion Recognition System

Dalleart et al. [136] produced the first practical work on automatic emotive identification in 1996. Five different emotions (sadness, anger, fear happiness, and neutral) were observed within the sample size of 1250 sentences. This comprised the database. With this information, a competed analysis was given, based on a variety of voice-related features and devices, show 80% recognition accuracy. All aspects and analyzed techniques were based solely on a pitch, built on underlying data that was very idealistic. Notably, the techniques used in the study bear a striking resemblance to those used today.

This forms the foundation of speech emotion recognition systems, a construct that identifies patterns that consist of a tripartite of major parts. Figure 13 provides a demonstration of these: signal processing, feature calculation, and classification. Another process follows the digitization of a recorded signal, the categorization of that signal, and the acoustic pre-processing of the material. This is known as signal processing, an endeavor which contains the steps above, all of which vary in importance and difficulty.

Feature calculation works toward the goal of locating the properties, contained within the digitized signal, that are of emotive importance and then representing them on an n-dimensional vector. Thus far, the features within such a programmed have not been fully classified as to their place within a ranking of importance, but a general agreement does exist on which facets of the endeavor are of most importance. Good
features seem to be highly data dependent, a field of study for Devilliers et al. [137]. The majority of attempts analyze a large number of characteristics, to which an algorithm is assigned so that the dimensionality of the pre-analysis data may be minimized. Furthermore, the algorithm selects the most important characteristics, concerning the specific task. Other methods include making use of principal components analysis so that the main data may be encoded in a more compact format.

Emotions are constructed in the form of a vector post the feature calculation analysis, leaving the enigma of identifying the emotion as a simple matter of recognizing patterns. This stretches to static and dynamic sequences. Static modeling is defined by one vector that represents one emotive unit. Dynamic modeling consists of multiple sequences of vectors to represent a single emotive unit, as well making reading for the temporary nature of emotional characteristics. With the complexities of this, the result is usually encoded temporary data within the vector. The dynamic model also features smaller vectors, making both options viable. The most common forms of these are Support Vector Machines, Bayesian classifiers, HMMs, and Neural Networks. Static modeling approaches are the current choice in this field.
A complete and thorough consideration of all the aspects of algorithms must be achieved for all real-time emotional recognition. Furthermore, the approaches require a certain amount of evolution to the environment, as any influences are not subject to alteration after that.

Correct procedure should see a final mention, concerning offline variation methods, administered to classifiers for the sake of pattern identification and emotion recognition before expounding on further details. For the accurate predictions of the success rate of a trained classifier using new data, it must be operated with information that it has not yet been subject to. Databases are usually a taxing venture; thus, existing databases are often used in the training of technology. However, the quality of the training dramatically increases with the access of new data, an amount that is usually not available. The results of such tests depend primarily on how the
chosen data is split. Another method is to split the data into any number of segments, taking n – 1 segments and using it for training, while the remainder comprises the data that is implemented in testing. Therefore, a greater overall accuracy occurs, a result of the accuracy on all test splits. N often equals 10; this paper makes use of n in the form of n = 3 or n = 5. Splits are proportioned in equitable sizes; class distribution, too, remains equal. This is namely “stratified cross-validation.” For the sake of quality in predictions, all criteria must be taken into account. There is a wide range of variables in the pursuit of classifying emotions. Therefore, things such as age, gender, independence, and personalities must be observed. Such variables must be distributed equally amongst splits.

3.4.2. Feature Calculation

There are signals that are used to interpret the emotions at play in any given sentence. They are given by a speaker and form a signal that is in the form of a periodic waveform. These waveforms are subject to the normal rules of any other such thin: frequencies of amplitude time, etc. All of these characteristics and their nuances provide information on the emotion. Figure 13 provides a graphical solution of this, with time on the x-axis and amplitude on the y-axis – frequency may then be deduced by oscillation.

A digitized waveform is sampled. The rate of this sampling is key, as the Nyquist theorem states that a loss of information need not be suffered provided frequency components do not exceed more than half the rate of sampling. Since the frequency of speech remains below 8kHz, a sampling rate of 16kHz suffices in order to provide information that is correct and complete. Telephones usually submit signals at 8kHz, making the phenomes seem more equalized than they actually are. However, human beings can tell differences that are inferred by context.

A spectrogram, also shown in figure 14, plays another role in the aid of understanding signals. It allows the representation of energy derived from frequency bands from
short time periods. Thus, it may be simply put as a representation of frequency over time. It is represented by the color. If a frequency is colored in a dark fashion, it means that the frequency contains lots of energy; light colors denote low energy. Speaking styles contain different frequency patterns, each is distinct and picked up by the spectrogram.

![Figure 13: Representation of a signal as a waveform and as a spectrogram.](image)

There is much to be derived from the study of amplitude, frequency and time, such as spectral properties like that which is shown in table 5. This may appear simple to derive, a simple observation for prosodic organizations, but this is not the case. There are a variety of tasks that are easy for humans that are not so for computers, as the multiple variables that the human brain instantly recognizes have not all been precisely defined. Therefore, we are not yet able to compute these in a manner that is conducive to human-like recognition. Studies like that of Murray and Arnott [138] show these contradictions to be true, as seen in anger – for anger has different types. The acoustic effects of this remain similar, but the emotion does vary. Such variations are easily detectable by humans but not by machines. Further arguments may come
about from the fact that it has not been disclosed whether the data contains samples from actors. Such sampling may be questioned with regards to its emotional authenticity, as an actor cannot surely present the same emotion as that of a real reflex to stimuli. Further questions are then asked as to whether the data from more “natural” sources are better than the data that is a product of acting. However, this will be discussed later in this paper.

All this sum to show that there is no easy method to go about extracting the underlying emotions from acoustics. The following will attempt to explain the measures used for the automatic recognition of emotion from acoustic means. The human method is fairly simple and well documented, but not completely satisfied. The temporal structures are categorized into two strata: short, fixed-length intervals and suprasegmental.

Pitch

Pitch is defined as the ear’s processing and response to the height of frequency [139], mostly just fundamental frequency f0. However, these two are not interchangeable and not altogether identical, as f0 is intrinsic to the scientific explaining of an acoustic wave; pitch is simply the perception of the ear. It does not take a genius to understand the concepts of pitch, but it is one of the most important property for determining emotion within speech. As stated in previous sections, the intensity of pitch is a sure giveaway as to the emotion. Boredom is insinuated by a neutral pitch; happiness by a high pitch; sadness by lower pitches. Therefore, it is undeniable that pitch plays a dramatic role in the expression of emotion, but it is not as critical to identification as most people believe. Usually, a pitch is an indicator in general terms.

Pitch may be calculated, for example, by the number of zero-crossings in a given domain. However, this method may be better tailored to the realm of music, as speech requires a different method of calculation. Speech frequency must be calculated by
looking at the maxima. The maxima of the autocorrelation frequency will show jumps in pitch and transitional features.

Figure 14: Emotion relevant acoustic properties shown for a neutral, a happy and a bored utterance of an actress taken from the Berlin Database of Emotional Speech [31]. The text spoken in each of the utterances was “Das will sie am Mittwoch abgeben.” (“She will hand it in on Wednesday.”).
Formants

Formants may be described as the local maxima within a given frequency range that is usually a result of resonance created in speech [139]. The global maxima are the fundamental frequency, with further local maxima being further formants f1, f2, etc. Figure 15 (b) provides a representation of these frequencies in relation to vowel sounds, and further explanation may be found by Biesack and Kempe [140], and Waaramaa et al. [141]. Both advance the notion that certain higher formants represent positive emotions. So too does Goudbeck et al. [142] demonstrate that there is a relationship between high arousal and higher formants in vowels – also a positive valence and a high second format. Negative emotions have been linked to less articulation by the work of Kienast et al. [143], again proving that a relationship does exist between formants and emotion.

Loudness

Loudness, although difficult to measure and thus referred to as a related feature, is most commonly viewed as the strength of sound as perceived by the ear. A Fourier transformation allows a representation and some sort of calculation of the signal. However, this may not be an accurate measurement of true loudness, as all noise adds to the strength of a signal. High energy is correlated with high arousal, a factor that depends on a variety of variables. Speaking style, phenomes, utterance type and the effective state of the speaker all play a role in the reading on the energy curve. The energy curve encodes information that may be deduced in order to shed light on the effective state of the human. For example, as can be seen in Figure 15 (c), variations of energy may be lower for bored speech.
Mel-frequency Cepstral Coefficients

MFCCS, or mel-frequency coefficients [145], have been used as a successful representation in a variety of areas: automatic speech recognition, speaker identification, and emotion identification. These are calculated using a filter and scale system which act to source a Fourier transformation. After this, a DCT acts to change the logarithm within a range into a cepstrum. An MFCCS is then created, with an amplitude of cepstrum. It is normal practice that only the first dozen coefficients are used. The net result of the process is that a system that is akin to human perception is created. It has the unique and human-like quality of filtering out all unnecessary information, and it further brings a good result in the goal of emotional identification [146, 147].

Therefore, one must realize that the correlation between speech and the underlying emotional state is not as close as it is often perceived. Thus, traditional methods are not capable enough for such identification. Above, a visual description of MFCCs in Figure 15 (d).

![Figure 15: An example of decomposition into frequency scales by wavelet transformation.](image)

60
Wavelets

Similar to Fourier transformations, which describe a signal in the form of other functions, wavelets provide the cosine and sine functions for Fourier transformation. They provide a square-integration as a basis for transformation whose result is localized in both frequency and time. This is done by creating changed and scaled versions of a signal so that a single representation (in both scale and time) is obtained.

This is shown in 16, a demonstration of wavelet transformations concerning speech. A critical factor is the time abilities that this method gives, as it plays important roles in recognition of emotions. However, this is still a novelty which is used but not often [149].

Other Spectral Features

There are other techniques used in the efforts of emotional recognition besides those that have been stated in the sections above. Such measures include: the spectral slope, mean, center of gravity and the use of distribution in the spectrum all reflect aspects of the emotional arousal. An illustration is found in the graphical representations of Figure 17.

![Neutral, Happy, Boredom Spectra](image)

*Figure 16: The frequency spectrum of an /a/ in neutral, happy and boring speech.*
3.4.3. Varying Duration and Speaking Rate

The concept of time plays a factor in the deduction of emotions. Variances such as the rate of speech, the length of diction and the length of a syllable are all factors that must be considered under the banner of timing. Furthermore, these occurrences are highly quantifiable – by computers as well as by human perception. There are gross measures that may be obtained by calculating the distance between the maxima of local energy. These methods may be further refined by combining various gross estimations [150].

Patterns within the speech rate may also aid the determining of the affective states. Rhythm pattern from vowel duration and the differences in pause durations may also encode information on the emotional states of the speaker [Luengo et al., 151]. Pauses may either indicate pondering or a fragmented speech style.

3.4.4. Voice Quality

Voice quality relates to the styles of speech such as harshness, whispery, screechy, etc. These manners of speaking allow for an intuitive estimation of the emotional state driving the style of dialogue. Some methods may assess the voice quality. The most common of these are voice breaks and jitter and shimmer performance. The latter usually study the states of pathological voices, but they may also provide needed analysis on emotional states within rhetoric. Voice breaks measured by the number of silenced sections in usually voiced segments; Jitter and shimmer analyses the variable of length and amplitude within pulses, as seen in Figure 18.
Such a study may deduce the pureness of voice by relating the energy of harmonic segments to the energy of the corresponding part. Voice quality is undoubtedly a method of obtaining emotional aspects of speech, but these methods are not usually a part of traditional study methods.

3.4.5. Applications

Such use of technology is highly applicable to a variety of scenarios. Such an environment where it may be successfully implemented is within the industry of call centers [152]. Applications in this sector allow readings of callers, allowing feedback for management purposes. Furthermore, the messages stored within voicemails may be ordered according to emotional states, allowing the most urgent needs to be transferred to the correct personnel within a shorter frame of time.

Further applications may be to reduce the stress levels of users or helping in the field of medicine through an analysis of psychological states. Lie detection may also be in need of such analysis, as insight into the psychological states may help in crime solving spheres. Such technology may also be integrated into the household setting, helping individuals adapt to one another’s emotional state. For example, gaming may use this as a way to deepen gaming experiences by altering content according to identified emotions.
3.4.6. Challenges

Emotions are a highly subjective matter, as humans are unique and diverse. Therefore, it is often difficult for humans themselves to identify the emotional states of their fellow man, a result being that computing such information becomes even more difficult. The magnitude of the range and combinations of emotions is almost infinite, as counting the many permutations may prove inconceivable. Furthermore, certain emotions are subtler than others and may not be contained within the voice. Current emotional systems can only fathom and contain a select number of emotions – most hold six. Further questions arise as to what data is considered the best in the field. As it is a subjective endeavor, there is a decision to be had concerning which data best demonstrates the current emotive state in a specific scenario. Thus far, researchers have not solved this dilemma. Lastly, there is a problem with regards to a set with which one may compare results. There is no standard version of a data set, and the debate on how to collect such data is extensive. There is a gap between the theoretical world where experiments take place and the success rate of real-world applications of the same technology. This has led to the introduction of bodies that seek to provide such standardized information for emotional identification [153]

3.5. Related Works

Related works are a result of the long tradition emotion recognition has in the field of psychology. This section will make mention of several works that, because of their excellence, deserve mention in a summarized explanation below.

The last few decades have seen some researchers in this field nominated for their work. The historical birth of emotion recognition belongs to the work of Dallaert, Polzin, and Waibel, as they compiled a recognition method that made use of statistical patterns. These patterns would then categorize speech according to the emotions encoded within. In this effort, they amassed over 1000 segments of
dialogue from a variety of speakers. These utterances contained four different emotive features: happiness, fear, anger, and sadness. Only the pitch was used for the classification. Thus, a novel technique was adopted, achieving 80% accuracy. A result of the book Affective Computing [89] has been that the broader community has seen that technology may aid the research of emotional intelligence. Therefore, we have an indicator of the trends [154]. Furthermore, Vogt [155] proves to be a competent dissertation on the matter of real-time recognition of emotions underlying speech. This work sees the introduction of systems that can classify emotions in a manner that effectively achieves the feat of real-time recognition using the EMO VOICE.

Feature sets are crucial for classifiers because they provide necessary training. These features vary, local; global and linguistic; acoustic. Studies such as that done by Schuller et al. [156] have made use of the feature selection method [157]. This particular study by Schuller et al. makes use of a Forward Floating System which shows the acoustic features ranked by the method. It then further indicates a combination of multiple features to produce more accurate results and recognition [158]. These were compared with the sets of features, number 1000, from the pitch energy and MFCC time series. These were found to be dramatically different from those within different types of datasets [159]; thus, a combination of results was integrated to obtain better emotion identification within databases.

Another noteworthy study concerning speech recognition [157] gave excellent documentation and explanation of the last 15 years. It provides an analysis of where the current literature is placed and the history that led to current research. In this, the authors state that the most challenging aspect of the future is still the matter of finding better data. The new features that are being included in current studies include gender and age. Thus, classifiers that have taken emotion into account perform better than those which do not [160]. This was recorded in a paper that stated the gender-
dependent emotional recognition systems are used with 90% accuracy. Thus, by considering gender in the equation, overall success rates may improve by 2%-4%.

3.6. Factors of Speech Emotion Recognition

There are three critical procedures that must be followed, according to the current literature on speech emotion identification. These features are signal processing, feature selection, and classification [161]. The first step involves the breakdown of whole signals into smaller segments. Second, features must be identified and then pulled out of these segments to show the varying emotions. Lastly, there must be labeling of models that are trained with specific emotions so that the prediction of new data may be more accurate. Following under this, one will see concepts of primary signal expounded. Thereby, these three steps will be explained in greater detail.

Basic Signal Concepts

The first step (signal processing) involves the digitization and computer representations of a signal, containing sampling and filtering. This section will contain critical concepts that will be defined as follows:

- **Frequency** – The amount of the same occurrences, which is measured in hertz (Hz).
- **Loudness** – Related to the square of the amplitude; a measure of the energy of the signal.
- **Amplitude** – The vertical axis value is about air pressure.
- **Linear Predictive Curve** – A conventional technique used to encode a signal. It provides simplification by smoothing out spectra.
- **Spectra** – A waveform consisting of other waveforms of different frequencies. The Fourier Transform can calculate it.
- **Linear Predictive Coding (LPC)**. LPC is a favorite technique to encode a speech signal. It can smooth out spectra and make finding the spectral peaks easily.
- **Spectogram** – This is a visual illustration of a broad range of frequencies contained within one sound. Furthermore, it can represent frequency over a period, as seen in Figure 18 where the x-axis pertains to time and the y-axis represents amplitude.

- **Formants** – A peak contained within the spectrum of sound is defined as a formant. It models verbal communication, with particular accuracy in lower forms.

![Figure 18: Analog signals of a recorded speech.](image)

**Speech Data Processing**

The process of data processing is the next step. It is an integral part of the process, as it allows for an analog signal to be digitized so that it is easily worked within a computational context.
The process starts as the signal is sampled. Next, the amplitude is studied using measuring devices at fixed timed intervals. This will be done in full obedience to the Nyquist Theorem, which requires the measuring speed to be of a certain standard. Following this, the value of the amplitude will be used as an integer, allowing for more compact storage of data. Lastly, the signal is transformed into a digital version so that it may be worked with by computation technology. This entire process is demonstrated in Figure 19.
Table 6: Different emotion units of speech.

<table>
<thead>
<tr>
<th>Unit</th>
<th>Features based on</th>
</tr>
</thead>
<tbody>
<tr>
<td>phonemes</td>
<td>Phonemes</td>
</tr>
<tr>
<td>Words</td>
<td>Words</td>
</tr>
<tr>
<td></td>
<td>Words in context</td>
</tr>
<tr>
<td></td>
<td>All phonemes</td>
</tr>
<tr>
<td></td>
<td>Vowels</td>
</tr>
<tr>
<td></td>
<td>Voiced consonants</td>
</tr>
<tr>
<td>Utterances and turns</td>
<td>Utterances</td>
</tr>
<tr>
<td></td>
<td>All words</td>
</tr>
<tr>
<td></td>
<td>Central words</td>
</tr>
<tr>
<td>Fixed length intervals</td>
<td>Fixed length intervals</td>
</tr>
<tr>
<td>Relative length intervals</td>
<td>Relative length intervals</td>
</tr>
</tbody>
</table>

Emotional representation requires the breakdown of any given signal into smaller segments, as larger units are not easily calculated with high accuracy. Thus, the smaller segments are used for accurate handling of data. However, the segments still need to be large enough to ensure that the acoustic properties of the emotion remain intact [161]. These technical splices in the signal need to follow specific methods, as seen in [157]. They may be cut according to words or frame/time. No regulations are surrounding this, and no best method may be prescribed as a general rule of thumb. However, the word-based splits have proven time-consuming and costly [162]. More information may be seen in table 6 [155], where different methods are used with advantages and disadvantages resulting from each of them.
Feature Extraction and Selection

There is a significant part of the process of emotion recognition. It is feature extraction, a method that does not have standardized regulations to follow and the feature sets used may also prove to be a matter of preference. This leads to different features adopted for various studies. However, the most common of these remain energy and pitch. All elements may be put under two broad classifications: acoustic or linguistic. For the sake of this paper, the matters concerning acoustic elements will be used. Acoustic features may be further classified as either a global feature or a short-term feature [155] – also identified as functionals and low-level descriptors, respectively [156]. These two different features are computed in different ways, with the short-term intervals measured in very short intervals and a constant distance; there is a summarization of the short-term features within Table 7. The information within short-term features is emotive information that is encoded within the timing, which can be used in conjunction with dynamic classifiers.

Global features (also known as functionals) are measured with statistical methods. Statistical functions like mean, mean, percentiles, etc are implemented as there is no timing data that may be extracted. However, these global features are not time sensitive, allowing for the use of Support Vector Machines in attempts of classification. For more detailed information on these methods, and classification thereof, see [157]
Table 7: A summary of LLDs/short-term features.

<table>
<thead>
<tr>
<th>Type of feature</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>Temporal aspects, for example F0 value (measured in milliseconds)</td>
</tr>
<tr>
<td>Energy</td>
<td>Describe the intensity or amplitude</td>
</tr>
<tr>
<td>Pitch</td>
<td>Related to the tone</td>
</tr>
<tr>
<td>Spectrum</td>
<td>Good representation of human characteristics, including formants, spectral slope, mean, center gravity, etc.</td>
</tr>
<tr>
<td>Cepstrum</td>
<td>Hold frequency as well as timing information. The most popular is MFCC</td>
</tr>
<tr>
<td>Voice quality</td>
<td>Includes jitter, shimmer, and Harmonic to Noise Ratio (HNR). Nonlinear operators such as Teager Energy Operator</td>
</tr>
<tr>
<td>Wavelet</td>
<td>Similar to MFCC but more localized. Contain more information</td>
</tr>
</tbody>
</table>

After this process of extracting the desired features, there is another process of selecting the correct features by the objectives of the study. In light of this, certain selection criteria were introduced like the Sequential Forward Selection (SFS) or the Sequential Floating Search (SFFS). These algorithms form a small proportion of the selecting process systems, all of which must be used with other technologies (Support Vector Machine for example) to arrive at the desired features [163, 159, 154]. Further reductions in the number of features in the scope of study may be done by the Principal Component Analysis [164, 163]. Another such method can be found in the use of the Enhanced Lipschitz Embedding (ELE). This method provides reduction through a geodesic distance evaluation [165]. Thus, the hope is that the reduction in features may prove to increase the end accuracy of the research, aiding
emotional recognition. This is a hope that has been fulfilled, as these methods have proved to increase accuracy and to save time and improve efficiency.

Classification

Classification is primarily seen as having two segments: classifiers that are of a dynamic nature and classifiers that are static. As alluded to in the previous section, logic would mean that the dynamic classifiers are used in conjunction with short-term LLDs. In contrast, the classifiers that are dynamic are used with global features that have been analyzed using statistical functions. Further factors in the decision process of which classifier to use include the dimensions of the data and the amount of training data that the classifier has been exposed to.

The realm of static classifiers provides housing to a variety of renowned algorithms. Below is a list of the algorithms that have achieved a status of reliability and fame:

- Maximum likelihood Bayes [166]
- Neural networks [167]
- Decision trees [168, 169]
- Bayesian classifiers [170, 171, 172, 173, 174]
- Gaussian Mixture Models [175, 172, 176, 177, 178]
- Linear Discriminant Analysis [179, 163, 176]
- k-nearest Neighbours [180]
- SVMs [181, 182, 183, 184, 185]
- Random Forest (RF)

Only a few of the classifiers within this list are sufficient to act within the field of emotional identification. Large volume data requires specific technology to provide desired accuracy and comparison. One of these is software which mines data: Weka [186] – this has been used commonly to navigate experiments and research. [187] has attempted to show the comparison of the performance of certain algorithms, the results of which demonstrate k-NNs to be better than other classifiers. Although the
performance of k-NNs is exemplary, it does not have the capabilities to perform well with mass. High-dimensional dimensional data renders the classifier insufficient. In this regard, SVMs [181, 182, 159] have been proven better, as wide research has shown. This research also shows that SVMs have the highest rate of correct identification, better than Decision Trees and Native boys – even better than K-NNs. As a result, SVMs have become the most widely used classifier within the research field of emotional recognition.

Hidden Markov Models [149, 188, 146, 189] have become the most commonly used algorithm in the category of dynamic classifiers. Amongst HMMs, SVM, and LDA [188, 189], HMMs have shown the best results in the classification of small segments of speech. Following this, research has gone into discovering the best designs of HMMs. The best designs have proven to be database and task reliant, also those that are used on smaller segments rather than larger ones. The implementation has been successful, and HMMs has found its place in Automatic Speech Recognition.

More recent work has seen the introduction to the usage of a hybrid of HMM and SVM [190]. HMM has served to provide exemplary work in system training and the SVM classifier created the means for the efficiency of system testing. Such a model has shown an improvement of 4\% when compared to a regular SVM classifier. As of today, classifiers of static nature enjoy more practical uses than their dynamic counterparts, but the debate has still not been settled. The issue of whether to use static or dynamic classifier remains a grey area.

### 3.7. Databases with Emotional Speech

Databases have shown to be important in this area of research. However, certain databases that contain select information of a select nature are more important than others. One such example is a database that contains emotional speech. This material is essential in the development of statistical data, as there is a recognition that the specifics of emotion may only be obtained from a data sample of sufficient
magnitude. Having stated that, the quality of the data is also crucial, as the size of the database is not the only factor. The quality of the recordings is an important factor for the training of classifiers, as it must have access to new information and data sets to learn. A more specific recording is preferred. The use of these allows for the classifier to better identify unseen data in the future. However, the size of the sample is required for the allowance of recognition in a variety of examples. As a result, the research has been naturally restricted to the information that is contained within databases. Therefore, the better the databases, the better the operation, effectiveness, and efficiency of classifiers.

Figure 20: PCA: a mapping of the original two dimensions onto the principal component axes.

There are a variety of uses for databases that contain emotional speech. Psychological studies would seem an apparent field of interest, as there is a rich and long tradition of emotional studies within the area. However, the uses of emotional speech databases are most useful to the emotional identification field of research. As per [191], some reviews were created discussing the topic of such databases. This paper will introduce, summarize, and briefly explain a few of them in the following section.
These databases are essential – that is not a debated point. However, the majority of emotional speech databases contain emotions that are not always natural. Mostly, these emotions stem from acting, inductive mechanisms, or perhaps spontaneity. Therefore, it is only natural that the use of such data is questioned with regards to the naturalness of the content. Despite this, the data that is extracted from acted emotions tend to provide the best results, as such databases contain strong, exaggerated levels of emotion. The Danish Speech corpus (DES) [192], along with the Surrey Audio Visual Expressed Emotion Speech (SAVVE) are amongst the most well-known acted databases internationally. These respective databases contain recordings of emotions, differing slightly in sample size and the amount of recorded emotion. SAVEE contains seven elementary emotional states which were obtained from 4 people. DES contains five such emotions which were derived from 4 people. This dissertation contains the SAVEE database as part of the making of a base model for our speech emotive identification workings.

Furthermore, there are two more famous databases worthy of mention. They contain data that is induced, prior to recording in a laboratory environment. Namely, these two databases are: the SmartKom corpus [194]; the German Aibo emotion corpus [195]. In regard to a more real baseline, call centre communication may be the answer. These are recordings of live dialogue between agents and the caller, providing for a database that is not subject to data that is, in some respects, unnatural [137].

The University of Southern California [196] is host to the Interactive Emotional Dyadic Capture database (IEMOCAP). This is a multi-modal multi-speaker database that was obtained through various researchers from the institution. It is freely available – all one has to do is visit their website where it is available to be downloaded [197]. The data contained within this amasses to around twelve hours. Furthermore, it contains a wide variety of sample types including audiovisual data, video, facial motion, transcriptions, and voice recordings. However, it is an acted
database, but all acting is done by professionals including the emotional expressions. A unique feature is the motion capturing, which allows for the recording of real emotional experience. The size of this makes the data highly valuable, particularly for research purposes. IEMOCAP allows for the modeling of multimodal emotive identification.

The three aforementioned databases have played a crucial role in the writing of the contents of this paper, as they provide the basis for the proof of this thesis. The first of these is Surrey Audio Visual Expressed Emotion. The University of Surrey and Ryerson University SMART Lab. Lastly, the University of Toronto. All of the above play a part in the seven fundamental emotional states: sadness, neutral, disgust, fear, anger, surprise, and happiness.

- Four male individuals (whose native language was English) were recorded by the Surrey Audio-Visual Expressed Emotion. These persons were all students whose age ranged from 27-31 [193]. Seven fundamental emotions were used in the construction of this database. Variables were controlled by the use of standardized sentences, as well as emotion-specific sentences. Therefore, not all the emotions were linked to the same dialogue.

- Two people participated in a study conducted by the University of Toronto. The recording was of one young female and one elderly man, all of which allowed for the recording of seven fundamental emotions. This study has the goal of finding the effect of emotion with regard to a variant of words.

- Obtaining data on the extreme versions of emotions was the goal of the Ryerson Audio-Visual Database of Emotional Speech and Song database. It is a derivative of the Ryerson University SMART Lab. To this end, twelve males and twelve females were recorded for seven fundamental emotions. All of the sample rhetoric used for testing were used in regard to these emotions.

Human-computer interaction tends to struggle with the emotions that tend to be more realistic. The more realistic and emotion, the less likely that effective recognition can
be achieved. Thus, figure 21 demonstrates the level of difficulty in emotive identification with different variations of databases.

Figure 21: Difficulty of speech emotion recognition with different types of databases.
CHAPTER 4: FRAMEWORK OF FACIAL AND SPEECH EMOTION RECOGNITION

In this chapter, explain the framework of merging both system facial expression and speech emotion recognition will be represented. As discussed in Chapter I, emotion recognition can be done via speech voice, speech content, facial expressions, as well as hand and body gestures. Currently, most of the approaches for emotion recognition mainly use one of these sources. In this dissertation, we tried to use the speech features and facial expression features to perform real time facial and speech emotion recognition application for smart phone using cloud computing. The speech information can be the audio file recorded from the built-in microphone and the video streams recorded from the video sensor, such as the build-in Webcam of a laptop or other video camera. Recorded audio and video files can also be used as the input of the system.

![Figure 22: Over View Structure of Real Time Facial Expression and Speech Emotion Recognition System.](image)

There are three major parts of the system: the facial expression recognizer, the speech emotion recognizer, and merging both system facial expression and speech recognizer. The core of this system is the merging of both system part in which the
emotion recognition results from the facial expression and speech emotion will be integrated together with feature-level fusion for audio-visual emotion recognition. After the resulting fusion, the final decision of the emotion will be given. Figure 22, show the general framework of fusion for audio-visual emotion recognition system design. The details of this work will be introduced separately in the subsection in this chapter.

4.1. Facial Expression Recognition

As described in Chapter 2, for facial expression recognition, usually four steps are needed: face detection, face tracking, feature extraction, and classification. The first one is mainly to process static images, while the second one is proposed for emotion recognition from videos or live stream from a smartphone camera.

Overview of the Facial Expression Recognition System

Through extensive study of different approaches to the problem of face action representation, appropriate algorithms were selected for each stage of a system.

![Image](image.png)

Figure 23: Overview structure of the real-time emotional face recognition face on mobile phone system.
The system is built in a traditional manner and consists of 4 stages: pre-processing (which has face detection and face tracking), feature extraction and classification (Figure 23). The system operates on image sequences taken from a video camera. Static images are used in training and testing procedures but the interaction with a system is designed for video analysis. This section includes the description of all three stages of the system. Algorithms used at each stage will be explained from a theoretical aspect. Next, the implementation details will be outlined and the system's behavior will be illustrated.

4.1.1. The Pre-processing

Emotional face recognition on mobile phones requires this stage to consist of two steps: firstly, acquiring a sequence of images (8 frames/second) using a video camera, secondly, detecting the facial region of the image and standardizing the properties for lighting the image. This stage is often not considered a main step in emotional face recognition but it is very important in obtaining accurate results as well as empowering the system to make it applicable to any data set of facial images.

Face Detection and Tracking

As can be assumed, detecting a face is simpler than recognizing the face of a specific person. In order to determine that a specific picture contains one (or more) faces, the general structure of a face should be defined. Luckily human faces do not greatly differ; noses, eyes, foreheads, chins and mouths are common features; all define the general appearance of a face.

Combining all the features together, something resembling a face is received. By determining if each of the characteristic features is similar to a part of our picture, it can be decided whether or not the picture contains a face. The match does not have to be accurate; it must just roughly match to the features in the image.

The image tracking mechanism used is Haar Cascades to detect objects from within the background. The tracking is much faster than the OpenCV Haar implementation.
To recognize a face from the side view or at an angle is usually a difficult job; such problems may require 3D Head Pose Estimation. In case the image is blur, subject is wearing glasses, image is not very bright or different parts of the image are brighter or darker even then the face recognition becomes a complex task [198]. Haar features are used in such classifiers which is shown below in figure 24.

![Haar Features](image)

**Figure 24: Haar features [199].**

4.1.2. **Feature Extraction by Landmark**

Facial landmarks are points on specific parts of the facial image which indicate, for instance, the location of the nose, the eyes, the brows, and the mouth within an image. These points are tracked to follow the facial muscles’ movements in time. If all facial landmarks are considered as a connected graph, we can assume that the density of the graph differs in each facial expression (e.g. the pre-trained facial landmark detector inside the Dlib library is used to estimate the location of 68 [x, y] coordinates that map the facial structures on the face). After the detection of facial landmarks, distinctive information is used to extract from landmarks with the help of “Center of Gravity (COG)” of all face landmarks, as shown in Fig. 25. Graphs are very useful mathematical tools that can provide a wealth of information regarding the interrelationships of spatial points, in this case, information concerning facial landmarks. In order to extract features from these facial landmarks, spectral graph
analysis is used, through which a characteristic vector, depicting areas of density in a graph, is extracted.

\[ X_{COG} = \frac{1}{68} \sum_{i=1}^{68} x_i \]  
\[ Y_{COG} = \frac{1}{68} \sum_{i=1}^{68} y_i \]

Where \( X_{COG} \) is the x-coordinate of COG and \( Y_{COG} \) is the y-coordinate of COG.

\[ x_{relative_i} = x_i - X_{COG} \]  
\[ y_{relative_i} = y_i - Y_{COG} \]  
\[ EUC_i = \sqrt{(x_i - X_{COG})^2 + (y_i - Y_{COG})^2} \]  
\[ \beta_{nose} = \tan^{-1}\left( \frac{y_{28} - y_{31}}{x_{28} - x_{31}} \right) \]  
\[ \theta_i = \tan^{-1}\left( \frac{y_i - Y_{COG}}{x_i - X_{COG}} \right) - \beta_{nose} \]

where \( i = 1,2,\ldots,68 \)

\[ F = \{ x_{relative_i}, y_{relative_i}, EUC_i, \theta_i \} i = 1^{68} \]  

Where \( EUC_i \) is the Euclidean distance between point \( i \) and the central point, \( \beta_{nose} \) is the angle of the nose, point-28 is the top of the nose, and point-31 is the tip of the nose. \( \theta_i \) defines the relative angle and is corrected when the face is not perfectly horizontal and the feature vector length is 272.
4.1.3. Classification by Support Vector Machines

The Support Vector Machine (SVM) is the successful and effective statistical classification machine learning approach. SVM is a linear classification that separates the classes in feature space by using hyper-panes. SVM constructs the optimal hyperpane which means finding the maximum margin between two classes. There are two main stages in SVM, training and testing stages. In the training stage, there are specific training data \((x_i, y_i)\) where \(i = 1, 2, \ldots, N\) (number of training data) and \(x_i\) is a vector space that will represent the feature image; \(y_i\) is the class type of \(x_i\) and its value is always 1 or -1.

SVM is a machine learning algorithm that is concerned with finding the optimal hyper-plane that separates the two classes in the feature space. The optimal hyperpane means finding the maximum margin between the two classes.
Firstly, the sample data training set that is given is represented as \( \{ x_i, y_i \} , i = 0,1,\ldots,m \). Where \( x_i \in R^n \) and \( y_i \in \{ +1, -1 \} \).

Support Vector Machine (SVM) has been widely used in various pattern recognition tasks. It is believed that SVM can achieve a near optimum separation among classes. In our study, we train SVMs to perform facial expression classification using the features we proposed. In general, SVM builds a hyperplane to separate the high dimensional space.

Support Vector Machines (SVMs) have been recently proposed by Vapnik and his co-workers [200] as a very effective method for general purpose pattern recognition. Intuitively, given a set of points belonging to two classes, an SVM finds the hyperplane that separates the largest possible fraction of points of the same class on the same side, while maximizing the distance from either class to the hyperplane. According to Vapnik [200], this hyperplane is called Optimal Separating Hyperplane (OSH) which minimizes the risk of misclassifying not only the examples in the training set but also the unseen examples of the test set.

Support vector machines (SVMs) [201] are originally designed for binary classification problem. How to effectively extend it for multi-class classification...
problem is still an on-going research issue. Several methods have been proposed where typically we construct a multi-class classifier by combining several binary classifiers. Some methods also have been proposed that consider all classes at once. As it is computationally more expensive to solve multi-class problems, comparisons of these methods using large-scale problems have not been seriously conducted yet.

The support vector machine [202], given labelled training data:

\[ D = \{(x_i, y_i)\}_{i=1}^l, x_i \in X \subset \mathbb{R}^d, y_i \in Y = \{-1,+1\} \]  

(9)

constructs a maximal margin linear classifier in a high dimensional feature space, \( \phi(x) \) defined by a positive definite kernel function, \( k(x,x') \), specifying an inner product in the feature space,

\[ \phi(x). (x') = k(x,x') \]  

(10)

A common kernel is the Gaussian radial basis function (RBF)

\[ k_{\text{RBF}}(x,x') = e^{-\gamma \|x-x\|^2} \]  

(11)

Where \( \gamma > 0 \), kernel parameters to be decided by the users. Where Linear classifier is defined as

\[ h_{w,b}(x) = g(w^T x + b) \]  

(12)

\[ \text{Where } w \text{ is weights and } b \text{ is bias.} \]

The discriminant function implemented by a support vector machine is given by

\[ f(x) = \sum_{i=1}^l \alpha_i y_i k(x_i,x) + b \]  

(13)
To find the optimal coefficients, of this expansion, it is sufficient to maximize the functional,

$$\mathcal{W}(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j k(x_i, x_j)$$

Which can be considered here as a cost function for optimization.

The popular SVM guide [203] suggests the following setting to train a kernel classifier.
1. Scale each feature to an interval like [-1; +1].
2. Use the Gaussian kernel. Choose $\gamma$; that gives the highest cross validation (CV) accuracy.
3. Obtain the model $w$ using the selected $\gamma$.

### 4.1.4. Database Collection Methods

The data gathering division of this procedure tries to find the optimum way to recognize which facet of the expression provides the most excellent information resource, and subsequently which method of coding the information best facilitates extraction. For quite some time data gathering procedures have encountered many obstacles. But motionless images are considered to be the most straightforward platform for recognizing and gathering data because the areas of interest, demonstrating expression, are stationary. Gathering data from video series brings in further obstacles i.e., face recognition. Several kinds of face recognizing methods have been designed, as expression is dependent upon of head movement, and they are thus adapted for tracking faces in the video.

Facial occlusion often compromises facial feature database gathering methodologies. There are a lot of researches identifying the requirement to crop images and even discard complete pictures owing to this obstacle. Facial occlusions involve artifacts, i.e. spectacles, or a hairy face, or can be structural distortions (e.g., due to trauma).
This has caused a lot of researches to subject these still pictures to a process of pre-screening. This circumvents the need for the removal procedure. In the recent past, there was a tendency towards resolutely including these kinds of problem pictures; the rationale is towards building a truly effectual and globally representative database, and the screening process has to cope with these kinds of problems. Cohn and Kanade recommend that the subject filtration should not take place until after extraction of features but be addressed by the data collection and extraction process. They are also recognized and address that there are individual differences in all appearances and personalities and there is also another factor of expressiveness, which alludes to variations in the extent of changes in morphology, smoothness, and occurrence of deep expressions.

There is also another obstacle to data gathering procedures, and that is a variance of the facial texture of people from different ethnicities and age group. Young children have clearer, low textured skin, and they are also free from facial hair in even eyebrows or head. Ethnicity can make difficulties by structural differences can be seen in the face. The contrast between the sclera and iris is very much dissimilar in Northern Europeans in comparison with Asians. Lyons states that fear detection is considered as problematic for Japanese images and by removing the "fear" classification, accuracy is enhanced. Ekman mentioned that rudimentary emotions are displayed in approximately the similar way, while structural dissimilarity and cultural factors affect general expression.

As we appreciate that emotional displays are essential in the routine lives of human beings, the requirement for and significance of automatic emotion detection has been increasing coupled with the rapid evolution of human and computer interface software. The data used in this application is stored in the Cohn-Kanade database. In 2000, the Cohn-Kanade (CK) database was published with the intention to promote investigation into automatically recognizing person facial expressions. I have taken
about 636 images of different people with their mixed emotions and identify six types of emotions (Happy, Sad, Fear, Surprise, anger, Disgust) which will be compared with the actual images.

The selection of images from the database is as follows:

![Figure 27: Examples of six basic facial expressions from the CK+ database Source: CK+database (© J. Cohn).](image)

The numbers of images of each emotion that have been assembled to be used in this project are (Happy: 69, Sad: 28, Fear: 25, Disgust: 59, Neutral: 327, Surprise: 83, and Angry: 45). These images were taken with high frame rate cameras.

For testing this application used three different datasets, the first one was explained earlier above and the second is KDEF dataset records facial expression images from 140 amateur actors (70 males and 70 females) at five different viewing angles. All actors aged between 20-30 years old. They have no beards, no mustaches, no earrings, no eyeglasses, and mostly no visible make-up during phone sessions. For our testing study, we only consider frontal images, resulting in 980 facial expression images (Happy: 140, Sad: 140, Disgust:140, Angry:140, Neutral:140, Surprise: 140, Fear: 140). Figure 28, shows some example images from the KDEF Dataset.
Finally, JAFFE Databases contains 213 facial expression images of six basic facial expressions and the neutral expression. All facial expression is taken from 10 Japanese female models. We use all 213 facial expression images for our experiment study (Happy: 31, Sad: 31, Fear: 32, Neutral: 30, Surprise: 30, and Angry: 30).

4.1.5. Facial Expression Testing

The last portion of the project verifies the performance of the system. It evaluates the complete project by utilizing test and train database of images. The testing and verification of the complete project have been done on a module basis. After formulation of every module, verification process undergoes on MATLAB using the images. The performance is measured in the form of percentage. This percentage is calculated when complete input test images are compared against training images. For example, for and emotion ‘Happy,’ total 100 images are input to the system and the system correctly identifies 92 of them to be the images of happy emotion and rest identifies as non-happy emotion. Then the performance of the system is 92 %.
Similarly, the system is verified on all the emotions and performance has been recorded. These performances have been shown in the following tables with the results of each emotion in the percentage of detection.

**Feature Extraction Test**

Landmarks have emerged as an important feature descriptor. Its value lies in its computational complexity as well as its recognition rate in most cases. A variant of the Landmarks + COG descriptor named Landmarks + COG -64 (i.e., 64 bytes descriptor) has been implemented. It has been shown in [42,44] that a 64 bytes descriptor is sufficient to obtain very good matching results.

The input for feature extraction is the region of interest, i.e. the face. This region is converted to grayscale. Gaussian smoothing [42,44] is applied on the resulting image, i.e. it is filtered using a Gaussian low-pass filter with a variance of Gaussian kernel taken as 2 [42,44] and a window size of 8x8 [42].

The testing of Landmarks + COG descriptors work on the pixel level and is thus very sensitive to noise. By pre-smoothing the patch, this sensitivity can be reduced, increasing the stability and repeatability of the descriptors. MATLAB simulation result of filtrating is shown in Figure (30 – 31).
Figure 30: Example from CK+ dataset used to Extract the feature using MATLAB.

Figure 31: Histogram of Facial Landmarks descriptors using CK+ dataset.

For all the patches, we then plotted out the histogram values found in the Facial Landmarks descriptors; this was done to be able to reflect the feature values depicted in graph form. Depending on the requirements, the number of bins can vary.
Fig. 31 shows an example of a histogram. If there is a lack of the color blue in any of the ten bins, this signifies the absence of any element in the bin. A column marked at the height of 1 indicates for that specific bin, there is one element contained in the bin. A column marked at the height of 2 signifies that there are two elements contained in the bin, et al.

Test the system through the phone

Output of real time facial expressions recognition from the app.

Figure 32: Overview of testing facial expression through the app.
Figure 33: Column visualization of three detected emotions of “Surprise,” “Sadness” and “Happiness.”

Figure 34: Line with visualization marks of three detected emotions of “Anger,” “Natural” and “Happiness.”
Figures (32-35), shows simple of testing using a smart phone and represent the percentages of facial expression recognition that been detect in real time. where Navy blue, yellow, red, green, royal blue, black, and gray Pie chart represent, Natural, fear, disgust, angry, surprise, sad, and happy facial expression recognition respectively.
Testing summery

The final step in this part is to assess and evaluate the project performance; to measure how many of the requirements for emotional face recognition on mobile phones can be achieved. Testing has been continuously addressed from the early implementation stage until the final stage.

Firstly, the testing of each function is carried out individually. It is tested to ensure that the algorithm and each line code works correctly. Sometimes, a small sample of the image data is used to test the code and sometimes a different kind of data set is built. For example, a database of the numeric matrix is built to calculate easily the results of the tested function manually and to compare the target results with the code results.

Secondly, after completing a certain stage, the performance of that stage is tested. Furthermore, after integrating the system stages, the overall system performance was tested. In these phases, sometimes an implemented algorithm that is not useful for this project is discovered; then one is forced to return to the starting point. For example, the Support Vector Machine (SVM) algorithm is described three times in three different papers. All these attempts failed in our hands. The problem was with the huge number of multi-class features that need to be trained. To solve this problem, attention was turned to the Android platform tools and a MATLAB tool that can be used with the project data. The Android platform was used to program the application and the MATLAB tool used to test the application. Eventually, after many attempts, the optimal solution was found.

Finally, the entire proposed project was tested to guarantee that the project met the requirements and achieved its aim. This is the most significant step, which ensures that the study algorithm is efficient and powerful; that it could recognize emotional face expressions from a mobile phone. The result of this testing and Evaluation are discussed in the next section.
4.2. The Implement of Speech Emotion Recognition

As discussed in Chapter 3, there are three steps that are general needed for the speech motion recognizer which are data pre-processing, classification, feature extraction, and selection. Figure 36, describes the typical automatic emotion recognition from speech. Firstly, an emotional speech database required for testing and training purposes. Appropriate acoustic emotion units must be separated from the instances. Following this, the acoustic characteristics are extracted, and the most relevant features relating to emotion recognition are selected. Finally, a quick and precise algorithm is required to classify the training data and to build a base model. In the classification stage, the emotions of test instances or real-time instances are forecasted based on the training model.

![Figure 36: Typical steps for speech emotion recognition [155].](image)

4.2.1. Signal Processing

As presented in the past section, in the signal processing stage the analogy speech data will be changed over to digital data. At that point, the digital will be portioned
into suitable units. Even though there are numerous kinds of the units, fixed size edges are generally utilized as a part of numerous popular, accessible toolboxes as a result of their productivity. Figure 37, gives an extended comprehension of the signal processing.

Figure 37: Speech signal processing.

4.2.2. Feature Extraction

The objective of feature extraction is to represent any speech signal by a limited number of measures (or features) of the signal. This is because the total amount of the data in the acoustic signal is excessively difficult to process, and not all of the data is applicable for particular assignments. In current ASR frameworks, the approach of extraction has for the most part been used to discover a representation that is moderately steady for various cases of a similar discourse sound, in spite of contrasts in the speaker or natural attributes, while keeping the part that speaks to the message in the speech signal generally flawless. The primary feature extraction procedures are Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP) and Mel-Frequency Cepstral Coefficient (MFCC). Though, MFCC is the most utilized feature extraction method in a programmed speech recognition framework.

Mel-Frequency Cepstral Coefficient (MFCC)

MFCC is a standout amongst the most mainstream feature extraction methods utilized as a part of automated speech recognition and speech emotion recognition, whereby it depends on the frequency domain of Mel scale for human ear scale [204, 205 and 206]. MFCC depends on the known variety of the human ear's basic transfer
speeds with recurrence. Speech signal has been communicated in the Mel frequency scale, to catch the critical qualities of phonetic in speech. This scale has a linear frequency spacing under 1000 Hz and a logarithmic spacing over 1000 Hz [208]. Typical speech waveforms may differ now and again relying upon the physical state of the speakers' vocal cord. Instead of the speech waveforms themselves, MFCCs are less vulnerable to the said varieties. MFCC block diagram comprises of the following steps:
1. Pre-processing.
2. Framing.
3. Windowing.
4. Discrete Fourier Transformation (DFT).
5. Mel-Filter bank.

Framing is the first step of the MFCC algorithm. During this stage, the time's intervals for the feature extraction is determined. In general, a 10ms to 30ms frame length is selected for speech recognition. Overlapped framing is utilized for efficient information extraction between the two adjacent frames. This means that for example
a frame of 30ms is shifted 10ms to have a new frame, 20ms of the previous frame in included in the new one [207]. In the windowing stage, a window function is applied to the frame. The Hamming window is considered to the most used windowing technique for the processing of speech. It is defined by the following formula:

\[ w(n) = 0.54 - 0.46 \cos \left( \frac{2\pi n}{N-1} \right), \quad 0 \leq n \leq N - 1 \] (15)

Where: \( N \) is the length in fraction of the window and \( n \) is the frame index.

The Fast Fourier Transformation (FFT) is then applied to the window to have the frequency content of speech signal in the current frame. The frequencies are then filtered by a Mel-scale filter that imitates the varying resolution of the human ear with frequency filters spaced linearly at low frequencies and logarithmically at high frequencies and which is defined as:

\[ \text{Mel} (f) = 2595 \times \log \left( 1 + \frac{f}{700} \right) \] (16)

Where: \( f \) is the frequency in Hz [209].

The Discrete Cosine Transformation (DCT) is applied to the logarithm of the mel-scale filtered frequencies. The first \( N \) coefficients (usually 13) are selected as the feature vector representing the selected frame [210].

4.2.3. Classification using Support Vector Machine

Support Vector Machine (SVM) is a much-known effective approach for pattern recognition. Here, the basic concepts of the SVM will be presented briefly. In the SVM approach, the primary aim of an SVM classifier is obtaining a function \( f(x) \), which determines the decision boundary or hyperplane. This hyperplane optimally separates two classes of input data points. This hyperplane is shown in Figure 39.
SVM is based on the idea to transform the original input set to a high dimensional feature space by using a kernel function. In this kernel function input space consisting of input samples are converted into high dimensional feature space, and therefore the input samples become linearly separable. It is clearly explained by using an optimal separation hyperplane in Figure 39.

The main advantage of SVM is that it has limited training data and hence has very good classification performance. For linearly separable data points, classification is done by using the following formula [211],

\[
\langle w \cdot x \rangle + b_0 \geq 1, \forall y = 1 \quad (17)
\]

\[
\langle w \cdot x \rangle + b_0 \geq -1, \forall y = -1 \quad (18)
\]

Where \((x,y)\) is the pair of the training set. Here, \(x \in \mathbb{R}\) and \(y \in \{+1, -1\}\). \(\langle w \cdot x \rangle\) represent the inner product of \(w\) and \(x\) whereas \(b_0\) refers to the bias condition. SVM that employs both the linear kernel function and the Radial Basis Kernel (RBF) function [212] is used here. The linear Kernel function is given by the formula below,

\[
\text{Kernel} (x, y) = (x \cdot y) \quad (19)
\]
The following formula gives the radial basis kernel function,

\[
\text{Kernel} (x, y) = e^{-\frac{||x-y||^2}{2\sigma^2}}
\]  

(20)

The SVM classifier places the decision boundary by using maximal margin among all possible hyperplanes.

The Support Vector Machine (SVM) is widely used as a classifier for emotion recognition for classification and regression purpose. It performs classification by constructing an N-dimensional hyperplane that optimally separates data into categories. The classification is achieved by a linear or nonlinear separating surface in the input feature space of the dataset.

4.2.4. Database Collection

Typically, the emotional database within emotional recognition is applied towards the study of phonetics and acoustics, along with research and development in the area of emotion speech recognition systems. For the purpose of this research, both RML and SAVEE DB have been studied.

- **SAVEE - Surrey Audio-Visual Expressed Emotion Database**

The SAVEE databases conducted the recording of video, audio and audio videos of four male actors that were aged between 27-31. The actors used seven different types of emotions, and there were approximately 480 different British speeches that were chosen from the TIMIT database. The data samples approximated 44.1 kHz for audio samples, and 60 fps for video samples. The seven different emotions that were used for classifying the audio samples were: surprise- 60, fear-60, sadness- 60, neutral-120, anger- 60, happiness- 60, and disgust -60.

Only audio data was used for this experiment whereby all of the speakers were spontaneously grouped into test (30%) and training (70%) sets.
Ryerson Multimedia Research Laboratory (RML) Database

Ryerson Multimedia Research Laboratory (RML) also makes ongoing efforts to build multimodal databases related to emotion recognition. The RML emotion database is language and cultural background independent audiovisual emotion database [214]. The video samples were collected from eight human subjects, speaking six different languages and six basic human emotions are expressed. It contains 720 audiovisual emotional expression samples.

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) Database

A RAVDESS dataset was utilized in [217], [218] which is used to investigate the similarities and variances in the visual and acoustic signals of emotional singing and speaking. This included the visual recordings of approximately 24 performers whereby 12 of them were male, and the other 12 were female. Both sets of performers spoke and sang the same two sentences with varied emotions at both a strong and usual emotional intensity and were repeated two times. The speech recordings consisted of eight emotions which were fear, anger, sadness, happiness, calmness, and finally, neutral. The singing performances had three melodies which were used: one was for the neutral emotion, the second was for the positive emotions, and the third was for the negative emotions. The main difference between each of the melodies was the two middle notes. 350 participants took parts in evaluating the emotion content found in the data sentences; each of the utterances received ten rating from ten different participants. They were asked to identify the express emotion that the performer was using from a set list which comprised all of the target emotions, or alternatively assert that none of them were correct. For each of the utterances, the range of agreement ranged from 0 to 1 and was calculated. For
a total consensus to be established between the participating evaluators, the score that would have been used was 1. However, if there was no agreement between the participants, then the score that was used was 0. The participators had to additionally rate the emotional intensity and genuineness for all of the performances. Only those performances that had an average intensity were used so that the performers were able to avoid having to exaggerate.

Furthermore, the dataset was also decreased further so that it was able to address two key matters. Firth is that there are two emotions which are unique to speech and the second was that each performer only had one speaking data. As a result, these features introduced the dependences that occurred between the performers and the different emotion which could have potentially resulted in a bias in the outcomes.

Accordingly, we were able to eliminate the disgust and surprise utterances and also dropped the performer with only speaking data. This resulted in 1104 audio-visual utterances (2 domains × 23 performers’ × 6 emotions × 2 sentences × 2 repetitions). The target emotion that the performers followed was used as the ground truth. The reason for this is that the perceived emotion labels were not yet available. 0.69 was the average agreement rate of the target emotion, and 16.7% was the chance rate of agreement and classification accuracy.

4.2.5. Speech Emotional Testing

The last portion of the speech emotional part verifies the performance of the system. It evaluates the complete project by utilizing test and train database of wave files. The testing and verification of the complete part have been done on a module basis. After formulation of every module, verification process undergoes on MATLAB using the wave files. The performance is measured in the form of percentage. This percentage is calculated when complete input test files are compared against training files. For example, for and emotion ‘Happy,’ total 118 wave files are input to the system and the system correctly identifies 110 of them to be the wave files of happy
emotion and rest recognizes as non-happy emotion. Then the performance of the system is 95.65 %. Similarly, the system is verified on all the emotions and performance has been recorded. These performances have been shown in the following tables with the results of each emotion in the percentage of detection.

Feature Extraction Test

One of the most important stages in the Speech Emotion Recognition System is the selection of appropriate features that carry the data about emotions from the speech signals. Many researchers have demonstrated that speech energy, formant frequency, fundamental frequency, and Mel frequency cepstrum coefficients are viable parameters for identifying and distinguishing particular emotional states. The feature extraction is founded on segregating speech into frames or smaller intervals [6]. Information regarding speech emotion is held within the pitch signal due to the fact it depends on the vocal fold tension. The vocal fold vibrations are also known as the fundamental frequency. For emotional estimating, the next significant feature is energy. This is because there is a change in the speech signal energy when emotions differ. In automated speech recognition and speech emotion recognition, Mel-frequency Cepstral Coefficient (MFCC) is one of the most used spectral features available [7][8], [9]. It has numerous advantages like simple calculation, the better ability of distinction, and high robustness to noise. Here MFCC features are extracted from the Praat software [10] with window length 20ms and time step 10ms. Generally, the Hamming window is preferred because of its high-frequency resolution and good sidelobe suppression properties. First, the silence regions present in the database was removed based on the zero–crossing rate and also by thresholding the energy. The silence region does not contain any useful information and is hence removed. Human perception of hearing does not follow a linear scale, and hence MFCC follows the Mel scale [11] which is a frequency scaling having linear spacing.
below 1000Hz and logarithmic spacing above 1000Hz. The formula to compute the Mel frequency for any given frequency \( f \) in Hz is given below [12],

\[
Mel (f) = 2595 \times \log(1 + \frac{f}{700})
\]

(21)

The Mel scale filter bank has a triangular series of uniform overlapping filters with constant bandwidth equal to 100 and their center frequencies at 50. This is what is believed to occur in the human auditory system [13]. This corresponds to the spacing on the Mel frequency scale.

**Speech Emotional Recognition Testing Through MATLAB**

The simulation of the testing speech emotion recognition system is built using MATLAB R2016a software. The sampling frequency of the audio signal is 44.1 KHz. The frame size is 236 samples with 118 sample frames overlapping. 12 mel coefficients with 300Hz and 3700 Hz as low frequency and high-frequency resp. The 20 mel filter bank is used as shown in Figure (40 – 44).

![Figure 40: Example from RAVDESS dataset used to Extract the feature of Sadness using MATLAB.](image-url)
Figure 41: Example from RAVDESS dataset used to Extract the feature of Happiness using MATLAB.

Figure 42: Example from RAVDESS dataset used to Extract the feature of Fear using MATLAB.
Figure 43: Example from RAVDESS dataset used to Extract the feature of Disgust using MATLAB.

Figure 44: Example from RAVDESS dataset used to Extract the feature of Anger using MATLAB.
Speech Emotional Recognition Testing Through Mobile Application

Testing speech emotion using the application that can recognize the mood of a user through their voice, by their mood and classify this into the classifier from a smartphone; through cloud computing to compare the actual result that is taken in real time with the dataset that has been stored in the cloud:


1) Anger Emotion

Anger requires high energy to be expressed. The definition and meaning of anger are simple and extreme displeasure. In the case of anger, aggression increases in which the control parameter weakens. Anger is stated to have the highest energy and pitch level when compared with the emotions disgust, fear, happiness, and sadness. The widest observed pitch range and highest observed rate of pitch change are other findings of the emotion label anger when compared with different emotions. Besides this, a faster speech rate is observed in angry speeches.

Figure 45: Anger Speech Emotion through the app.
2) Disgust Emotion

Increased articulation precision at stressed content words was noted. The pitch contour showed downward inflections at the phrase endings, and also downward pitch inflections at word endings. A rise in pitch was noted at the beginning of stressed content words. The speech rate was low, with a large number of introduced pauses, increased phonation time, and lengthening of the stressed syllables in stressed content words.

Intensity was quite loud, which is not a typical characteristic for disgust, though it decreased towards the end of the utterance. For all of the disgust utterances, an increase in articulation precision was noted. Again, there was an emphasis on the pitch contour changes, with much accenting and use of high intensity. Large dynamic changes within the intensity contour were noted, and variations in contour between utterances. A low mean pitch level, a low-intensity level, and a slower speech rate are observed when disgust is compared with the neutral state. Disgust is stated as the lowest observed speech rate and increased pause length.
3) **Happiness Emotion**

An increase in articulation precision was noted for content words, and the voice generally sounded breathy. The general form of the pitch contour was with a second rising part towards the end of the utterance, with a terminal fall. It was noted that the line of the pitch contour was not smooth; it had sharp small oscillations at the primary stressed syllables and local downward pitch changes which seem to be rhythmic (stressed phonemes occurring at regular intervals).

Happiness exhibit a pattern with a high activation energy, and positive valence. The strength of the happiness emotion may vary. In the emotional state of happiness or joy, pitch means, range, and variance increases. It is stated that fundamental and formant frequencies increase in the case of a smile. Moreover, amplitude and duration also increase for some speakers.
4) Sadness Emotion

As the sad voice exhibited an overall decrease in articulation precision. Small downward inflections at the phoneme level were noted, and there were regular pauses. Small downward inflections were noted at word and phoneme level. A low-intensity contour was noted for all the utterances, with intensity decreasing towards the ends.

In an emotional dimension, sadness requires very low energy. Also, valence degree is negative. Sadness exhibits a pattern that is normal or lower than normal average pitch, a narrow pitch range and slow tempo. Speech rate of a sad person is lower than the neutral one.
Figure 48: Sadness Speech Emotion through the app.

5) **Neutral Emotion**

For both actors, the utterances spoken with neutral emotion are clearly articulated speech and show some pausing between words.
6) **Surprised Emotion**

Noted that with surprise “the voice suddenly glides up (or up-down), falls to a mid-level (joyful surprise) or a lower level (stupefaction). Also noted a very wide pitch range for a surprise, with tempo and pitch median normal or higher.
Testing summery

The final step in this part is to assess and evaluate the application performance; to measure how many of the requirements for emotional speech recognition on mobile phones can be achieved. Testing has been continuously addressed from the early implementation stage until the final stage.

Firstly, the testing of each function is carried out individually. It is tested to ensure that the algorithm and each line code works correctly. Sometimes, a small sample of the wave file is used to test the code and sometimes a different kind of data set is built. For example, a database of the numeric matrix is constructed to calculate the results of the tested function easily manually and to compare the target results with the code results.
Secondly, after completing a particular stage, the performance of that stage is tested. Furthermore, after integrating the system stages, the overall system performance was tested. In these phases, sometimes an implemented algorithm that is not useful for this research is discovered; then one is forced to return to the starting point. For example, the Support Vector Machine (SVM) algorithm is described three times in three different papers. All these attempts failed in our hands. The problem was with the considerable number of multi-class features that need to be trained. To solve this problem, attention was turned to the Android platform tools and a MATLAB tool that can be used with the project data. The Android platform was used to program the application and the MATLAB tool used to test the application. Eventually, after many attempts, the optimal solution was found. Finally, the entire proposed project was tested to guarantee that the project met the requirements and achieved its aim. This is the most significant step, which ensures that the study algorithm is efficient and robust; that it could recognize emotional speech from a mobile phone. The result of this testing and Evaluation are discussed in the next section.

4.3. Information Fusion

In the process of emotion recognition, information fusion refers to combining and integrating all incoming information into one representation of the emotion expressed by the user. There are two problems to solve in this step: when to integrate the information and how to integrate the information [215]. Generally, there are two primary information fusion methods: feature level fusion and decision level fusion.

Feature level fusion is done on the feature sets extracted from the speech and facial expression, and the process is shown in Figure 40. However, the feature level fusion cannot be normalized well when the information is different in the temporal characteristics. One requirement for the extracted features is that they should be
synchronous and compatible. Another problem to be figured out for feature level fusion is the high dimensional data which can result in massive computing and time-consuming.

Decision level information fusion is based on the assumption that different modules are independent of each other. In this kind of method, each module (in our case is speech emotion recognition or facial expression recognition) is classified separately, and the output of each module is integrated to get the global division of the expressed emotion (see Figure 51). Currently, the decision level information fusion method is widely used for multimodal emotion recognition.

4.3.1. Feature Level Fusion

Feature level fusion is a kind of solution which combines different feature vectors, obtained either with different modalities or by applying different feature extraction algorithms to the same modality. Figure 51, shows block diagram of feature-level fusion for audio-visual emotion recognition problem.

![Block diagram of Audio-Visual Emotion Recognition using Feature-Level Fusion.](image)

Feature-level fusion is concatenating two feature vectors. One of the feature vectors is obtained from audio data and the other one is obtained from visual data. Figure 52, illustrates the concatenating of two feature vectors for feature level fusion.
Feature-level fusion can be summarized with the following steps:

1. Obtaining raw data
   
   1.a. Audio data
   
   1.b. Visual data

2. Feature extraction
   
   2.a. Feature extraction from audio data
   
   2.b. Feature extraction from visual data

3. Concatenating of feature vectors

\[
\begin{align*}
    f_a(i), & \quad i = 1, 2, \ldots, N \\
    f_a & = \begin{bmatrix} f_{a1}^1, & f_{a2}^1, & \cdots, & f_{aN}^1 \end{bmatrix} \\
    f_v(k), & \quad k = 1, 2, \ldots, M \\
    f_v & = \begin{bmatrix} f_{v1}^1, & f_{v2}^1, & \cdots, & f_{vM}^1 \end{bmatrix} \\
    f_f(j), & \quad j = 1, 2, \ldots, N + M \\
    f_f & = \begin{bmatrix} f_{a1}^1, & f_{a2}^1, & \cdots, & f_{aN}^1, & f_{v1}^1, & f_{v2}^1, & \cdots, & f_{vM}^1 \end{bmatrix}
\end{align*}
\]
where $f_a$ represents feature vector of audio data which includes $N$ features, $f_v$ denotes feature vector of visual data which consists of $M$ features and fused feature vectors is showed with vector $o f f_r$. Length of fused feature vectors is equal to summation of the length of two feature vectors ($length(f_r) = N+M$).

There are many reasons to use decision level information fusion instead of feature level information fusion [216]. The feature level fusion needs to use large multimodal dataset because the feature sets are in a high dimensional data space. In the decision level fusion, multi-modules can be processed asynchronously. Besides, it provides more flexibility. Thus, we can use different classifiers on different data sets and integrate them without retraining. By using the decision level fusion, the recognizer can be used for single module emotion recognition. Moreover, it is more flexible and extendable using the decision level information fusion.

The difficulty for the information fusion is that the formats for speech emotion recognition and facial expression recognition are different. The speech emotion recognition is for each turn, and there are blanks when the turn stops or before the starting. But the results in facial emotion recognition are frame by frame. To fuse the results together, we have to know the frame rate of the video and then calculate the emotion.

### 4.3.2. Classification using Support Vector Machine

Support Vector Machine (SVM) is a much-known effective approach for pattern recognition. Here, the basic concepts of the SVM will be presented briefly. In the SVM approach, the main aim of an SVM classifier is obtaining a function $f(x)$, which determines the decision boundary or hyperplane. This hyperplane optimally separates two classes of input data points. This hyperplane is shown in Figure 35.
SVM is based on the idea to transform the original input set to a high dimensional feature space by using a kernel function. In this kernel function input space consisting of input samples are converted into high dimensional feature space, and therefore the input samples become linearly separable. It is clearly explained by using an optimal separation hyperplane in Figure 53.

The main advantage of SVM is that it has limited training data and hence has very good classification performance. For linearly separable data points, classification is done by using the following formula [211],

\[
\langle w \cdot x \rangle + b_0 \geq 1, \forall y = 1
\]  
(28)

\[
\langle w \cdot x \rangle + b_0 \geq -1, \forall y = -1
\]  
(29)

Where \((x, y)\) is the pair of the training set. Here, \(x \in \mathbb{R}\) and \(y \in \{+1, -1\}\). \(\langle w \cdot x \rangle\) represent the inner product of \(w\) and \(x\) whereas \(b_0\) refers to the bias condition. SVM that employs both the linear kernel function and the Radial Basis Kernel (RBF) function [212] is used here. The linear Kernel function is given by the formula below,

\[
\text{Kernel} (x, y) = (x \cdot y)
\]  
(30)
The following formula gives the radial basis kernel function,

\[
\text{Kernel } (x, y) = e^{-\frac{||x-y||^2}{2\sigma^2}} \quad (31)
\]

The SVM classifier places the decision boundary by using maximal margin among all possible hyper planes.

The Support Vector Machine (SVM) is widely used as a classifier for emotion recognition for classification and regression purpose. It performs classification by constructing an N-dimensional hyperplane that optimally separates data into categories. The classification is achieved by a linear or nonlinear separating surface in the input feature space of the dataset.

4.3.3. Database Collection

Typically, the emotional database within emotional recognition is applied towards the study of phonetics and acoustics, along with research and development in the area of information fusion recognition systems. For the purpose of this research, RML, SAVEE and RAVDESS DB have been studied.

- **SAVEE - Surrey Audio-Visual Expressed Emotion Database**

The SAVEE databases conducted the recording of video, audio and audio videos of four male actors that were aged between 27-31. The actors used seven different types of emotions, and there were approximately 480 different British speeches that were chosen from the TIMIT database. The data samples approximated 44.1 kHz for audio samples, and 60 fps for video samples. The seven different emotions that were used for classifying the audio samples were: surprise- 60, fear-60, sadness- 60, neutral-120, anger- 60, happiness- 60, and disgust -60.

Audio and video data were used for this experiment whereby all of the speakers were spontaneously grouped into test (30%) and training (70%) sets.
Ryerson Multimedia Research Laboratory (RML) Database

Ryerson Multimedia Research Laboratory (RML) also makes ongoing efforts to build multimodal databases related to emotion recognition. The RML emotion database is language and cultural background independent audiovisual emotion database [214]. The video samples were collected from eight human subjects, speaking six different languages and six basic human emotions are expressed. It contains 720 audiovisual emotional expression samples.

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) Database

A RAVDESS dataset was utilized in [217], [218] which is used to investigate the similarities and variances in the visual and acoustic signals of emotional singing and speaking. This included the visual recordings of approximately 24 performers whereby 12 of them were male, and the other 12 were female. Both sets of performers spoke and sang the same two sentences with varied emotions at both a strong and normal emotional intensity and were repeated two times. The speech recordings consisted of eight emotions which were fear, anger, sadness, happiness, calmness, and finally, neutral. The singing performances had three melodies which were used: one was for the neutral emotion, the second was for the positive emotions, and the third was for the negative emotions. The main difference between each of the melodies was the two middle notes. There were 350 participants who took parts in evaluating the emotion content found in the data sentences; each of the utterances received ten rating from ten different participants. They were asked to identify the express emotion that the performer was using from a set list which comprised all of the target emotions, or alternatively assert that none of them were correct. For each of the utterances, the range of agreement ranged from 0 to 1 and was
then calculated. For a total consensus to be established between the participating evaluators, the score that would have been used was 1. However, if there was no agreement between the participants, then the score that was used was 0. The participators had to additionally rate the emotion intensity and genuineness for all of the performances. Only those performances that had a normal intensity were used so that the performers were able to avoid having to exaggerate.

Furthermore, the dataset was also decreased further so that it was able to address two key matters. First is that there are two emotions which are unique to speech and the second was that each performer only had one speaking data. As a result, these features introduced the dependences that occurred between the performers and the different emotion which could have potentially resulted in a bias in the outcomes.

Accordingly, we were able to eliminate the disgust and surprise utterances and also dropped the performer with only speaking data. This resulted in 1104 audio-visual utterances (2 domains × 23 performers’ × 6 emotions × 2 sentences × 2 repetitions). The target emotion that the performers followed was used as the ground truth. The reason for this is that the perceived emotion labels were not yet available. 0.69 was the average agreement rate of the target emotion, and 16.7% was the chance rate of agreement and classification accuracy.
Test the system through the phone for the fusion option

Figure 54: Overview of testing fusion through the app.
Figures (54), shows simple of testing using a smart phone and represent the percentages of fusion (Facial expression and speech emotion recognition) that been detect in real time. where Navy blue, yellow, red, green, royal blue, black, and gray Pie chart represent, Natural, fear, disgust, angry, surprise, sad, and happy fusion option respectively.

4.3.4. Summary

In this chapter, the approaches for speech emotion recognition, facial expression recognition, and information fusion have been introduced in detail. The framework of the multi-sensory emotion recognition is shown in Figure 51. For the speech emotion recognition, there are three steps: speech processing, feature extraction, and selection, as well as the classification. All these works can be done with the android app. Development. To get a more stable base model for the speech emotion recognition, MATLAB was use to analyze the extracted features, accuracy, and other details of the results from three different emotional speech databases, RAVDESS, RML, and SAVEE. The results show that using 862 features with the SVM classification method can achieve better performance than using another feature set or classification method. There are two achievements in facial emotion recognition. The first approach is used to recognize facial expression from static images by the techniques of skin color segmentation. While the second one is based on the ASM face tracking and is perfect for recognizing facial emotion from live streams of a Webcam. No matter which achievement, there are some common steps, and they are face detection, feature extraction, and classification. Finally, as shown in Fig 51, the information fusion was done on a mobile phone using cloud computing. With the design of our multi-sensory emotion recognition, the emotion recognition accuracy has been improved by 1.83% compared to the speech emotion recognition and 0.96% compared to the facial emotion recognition. The details of our experiments and results are shown in Chapter 5.
CHAPTER 5: Experimental result and Analysis

So far, the framework of the multi-sensory emotion recognition has been introduced, including the facial expression recognition, speech emotion recognition, and information fusion. In this chapter, many experiments have been done based on the proposed work, and the results were examined in detail.

5.1. Choosing Base Model for Facial Expression Recognition

In the facial expression recognition approach, we have already built many different models based on different databases, feature sets, as well as the classification algorithm. According to the conclusion, the Cohn-Kanade, JAFFE and KDEF datasets based on Support Vector Machine (SVM) algorithm with 1829 images have been used for testing purposes.

The tests in this section were done with mobile application and MATLAB. As we know the Cohn-Kanade is English, in the JAFFE is Japanese, and KDEF is Swedish. The following tests will also show how the ethnicity affects the accuracy of the facial expression recognition.

Usage Data set for Evaluation

A set of 780 images has been used for testing purposes. This data set is a mixture of person images depicting happy, sad, angry, fear, disgust and surprised emotions. In this experiment, the Cohn-Kanade database has been used to examine the efficacy of the emotional face recognition system. Figure 5 shows some sample images from the database. This data base contains 130 images of each emotion, i.e. disgust, fear, sad, surprise, happy and angry, as described previously.
Tools used for Evaluation

The main and most useful statistical measurements that were utilized to evaluate the emotional face recognition on mobile phones are recall, precision, and accuracy. These measures are critical in that they are key elements in judging the performance of the emotional face recognition. The recall rate measures and studies the relationship between the correct classification rates of specific emotions and the wrong classification of this specific emotion whereas precision measures the relation between the correct classification rate of specific emotions and the wrong classification of other emotions that are classified as special emotions. Finally, the accuracy rate measures the relation between the correct classification rate of specific and other emotions and the total number of tested images. The following display these relations as symbolic equations.

\[
Recall = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \quad (32)
\]

\[
\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \quad (33)
\]
\[ Accuracy = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{false positive} + \text{true negative} + \text{false negative}} \] (34)

Where true positive for specific data set of emotions A is the correct classification rate of emotion; false positive of A emotion is the wrong classification rate of other data set emotions that are classified as type A whereas false negative is the wrong classification of emotion A. Finally, a true negative of emotion A is the correct classification for images whose label is not A. Additionally, it is noted that the summation of true positive, false positive, false negative and true negative is the total number of the test images.

Experimental Results

The performance of Facial landmarks and COG feature extracted and Support Vector Machine (SVM) algorithm has been examined. The experiment focuses on three facial expression datasets, namely Extended Cohn-Kanade Facial Expression Dataset, Japanese Female Facial Expression dataset, and Karolinska Directed Emotional Faces dataset. All faces in the three datasets were extracted with face detector of Computer Vision system in Android platform library. The system has been tested over the grayscale images. The previous source has been used separately.

To avoid confusion, firstly, the emotion recognition system has been tested using grayscale images. In this testing, the system has examined its ability to distinguish between seven types of emotion: natural, happiness, sad, angry, fear, disgust and surprise.

It is noticed that it has been done by using three different datasets. The following tables display the results that are collected from the three datasets files. Table 8, demonstrates the system results of (CK+) dataset which contains 636 images from 123 subjects, where 327 sequences are labeled with one of seven facial expressions. Table 9, demonstrates the system results of (JAFFE) dataset which contains 213 images of 7 facial expressions. Table 10, demonstrates the system results of (KDEF)
dataset which contains 980 images of 7 facial expressions. The formal that been used to calculate the accuracy for these three tables is:

\[ \text{Accuracy} = \frac{\text{correct predicted}}{\text{Total number of images}} \times 100\% \quad (35) \]

Example:

\[ \text{Accuracy} = \frac{55}{65} \times 100\% = 84.6\% \]

Table 8: Confusion matrix SVM classification and number in the 1th database of Cohn and Kanade (CK+) dataset.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>N</th>
<th>F</th>
<th>S</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral (N)</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Fear (F)</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Surprise(S)</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Happy (H)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Sad (S)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>87.5</td>
<td>0</td>
<td>12.5</td>
<td>100%</td>
</tr>
<tr>
<td>Angry (A)</td>
<td>0</td>
<td>5.88</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>92.3</td>
<td>7.6</td>
<td>100%</td>
</tr>
<tr>
<td>Disgust (D)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>94.1</td>
<td>5.9</td>
<td>100%</td>
</tr>
<tr>
<td>Average</td>
<td>96.3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Confusion matrix SVM classification and a number of images in the 2nd database of Japanese Female Facial Expression (JAFFE) dataset.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>N</th>
<th>F</th>
<th>S</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral (N)</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Fear (F)</td>
<td>11.1</td>
<td>88.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Surprise(S)</td>
<td>11.1</td>
<td>0</td>
<td>88.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Happy (H)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>88.9</td>
<td>11.1</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Sad (S)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Angry (A)</td>
<td>0</td>
<td>11.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>88.9</td>
<td>0</td>
<td>100%</td>
</tr>
<tr>
<td>Disgust (D)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12.5</td>
<td>0</td>
<td>87.5</td>
<td>100%</td>
</tr>
<tr>
<td>Average</td>
<td>93.03%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 10: Confusion matrix SVM classification and number of images in the 3rd database of The Karolinska Directed Emotional Faces (KDEF) dataset.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>True</th>
<th>N</th>
<th>F</th>
<th>Su</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral (N)</td>
<td>88.1</td>
<td>4.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.3</td>
<td>4.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fear (F)</td>
<td>0</td>
<td>90.5</td>
<td>4.7</td>
<td>2.3</td>
<td>0</td>
<td>2.3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surprise(S)</td>
<td>4.7</td>
<td>0</td>
<td>90.4</td>
<td>4.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy (H)</td>
<td>2.4</td>
<td>0</td>
<td>0</td>
<td>92.8</td>
<td>2.3</td>
<td>2.3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sad (S)</td>
<td>0</td>
<td>4.7</td>
<td>2.3</td>
<td>2.3</td>
<td>90.4</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angry (A)</td>
<td>2.3</td>
<td>2.3</td>
<td>0</td>
<td>2.3</td>
<td>0</td>
<td>92.8</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disgust (D)</td>
<td>0</td>
<td>2.3</td>
<td>4.7</td>
<td>0</td>
<td>2.3</td>
<td>0</td>
<td>90.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>90.8%</td>
</tr>
</tbody>
</table>

Note that the numbers of the test images for each emotion are not equal because the images for testing have been selected randomly from the test file. Moreover, the system may test an image more than once.

Moving to a more accurate result, the tables below illustrate the result of the system over the database file that contains whole image training and testing. The method that been used to calculate these three tables is Recall, Precision, and Accuracy. The system behavior over three datasets that have been examined and using 70% as training images with 30% as testing images for each dataset. The first dataset has 636 images, the second also has 213 images, and the third dataset has 980 images. The following tables demonstrate the system performance with each group of data.
Table 11: Confusion matrix using SVM classification for CK+ dataset.

<table>
<thead>
<tr>
<th>Predicted True</th>
<th>F</th>
<th>S</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>N</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>87.5%</td>
<td>100%</td>
<td>99.5%</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Happy</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Sad</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>100%</td>
<td>87.5%</td>
<td>95.5%</td>
</tr>
<tr>
<td>Angry</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>100%</td>
<td>92.3%</td>
<td>99.5%</td>
</tr>
<tr>
<td>Disgust</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>94.1%</td>
<td>94.1%</td>
<td>98.4%</td>
</tr>
<tr>
<td>Neutral</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>98</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>97.4%</td>
<td>96.3%</td>
<td>99.5%</td>
</tr>
</tbody>
</table>

Table 12: Confusion matrix using SVM classification for JAFFE dataset.

<table>
<thead>
<tr>
<th>Predicted True</th>
<th>F</th>
<th>S</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>N</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>88.8%</td>
<td>88.8%</td>
<td>96.6%</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>100%</td>
<td>88.8%</td>
<td>98.3%</td>
</tr>
<tr>
<td>Happy</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td>88.8%</td>
<td>98.2%</td>
</tr>
<tr>
<td>Sad</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>81.8%</td>
<td>100%</td>
<td>96.6%</td>
</tr>
<tr>
<td>Angry</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td>88.8%</td>
<td>98.3%</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>100%</td>
<td>87.5%</td>
<td>98.2%</td>
</tr>
<tr>
<td>Neutral</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>81.8%</td>
<td>100%</td>
<td>96.6%</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>93.2%</td>
<td>91.8%</td>
<td>97.9%</td>
</tr>
</tbody>
</table>
Table 13: Confusion matrix using SVM classification for KDEF dataset.

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear</td>
<td>86.4%</td>
<td>90.5%</td>
<td>96.4%</td>
</tr>
<tr>
<td>Surprise</td>
<td>88.4%</td>
<td>90.4%</td>
<td>96.7%</td>
</tr>
<tr>
<td>Happy</td>
<td>88.6%</td>
<td>92.8%</td>
<td>97.1%</td>
</tr>
<tr>
<td>Sad</td>
<td>95%</td>
<td>90.5%</td>
<td>97.8%</td>
</tr>
<tr>
<td>Angry</td>
<td>92.8%</td>
<td>92.8%</td>
<td>97.8%</td>
</tr>
<tr>
<td>Disgust</td>
<td>95%</td>
<td>90.5%</td>
<td>97.8%</td>
</tr>
<tr>
<td>Neutral</td>
<td>90.2%</td>
<td>88.1%</td>
<td>96.7%</td>
</tr>
<tr>
<td>Average</td>
<td>90.9%</td>
<td>90.8%</td>
<td>97.1%</td>
</tr>
</tbody>
</table>

Table 14: Total number of images collected for Facial Expression Recognition study.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Total Number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CK+</td>
</tr>
<tr>
<td>Name of</td>
<td>Train 70%</td>
</tr>
<tr>
<td>Dataset</td>
<td>Test 30%</td>
</tr>
<tr>
<td>Happy</td>
<td>49</td>
</tr>
<tr>
<td>Sad</td>
<td>20</td>
</tr>
<tr>
<td>Fear</td>
<td>18</td>
</tr>
<tr>
<td>Surprise</td>
<td>59</td>
</tr>
<tr>
<td>Anger</td>
<td>32</td>
</tr>
<tr>
<td>Disgust</td>
<td>42</td>
</tr>
<tr>
<td>Neutral</td>
<td>229</td>
</tr>
<tr>
<td>Total</td>
<td>636 images</td>
</tr>
</tbody>
</table>
5.2. Choosing Base Model for Speech Emotion Recognition

In the speech emotion recognition approach, we have already built many different models based on different databases, feature sets, as well as the classification algorithm. According to the conclusion, the RAVDESS, RML and SAVEE datasets based on Support Vector Machine (SVM) algorithm with 988 acoustic features were used. The tests in this section were done with mobile application and MATLAB. As we know the RAVDESS is English, in the RML is max of (English, Chinese, Urdu, Persian and Italian) and SAVEE is English. The following tests will also show how the language affects the accuracy of the speech emotional recognition.

Usage Dataset for Evaluation

A set of 2880 files has been used for testing purposes. This dataset is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. It contains 8 emotions such as Calm, neutral, happy, sad, angry, fear, disgust and surprised emotions. In this experiment, the RAVDESS database has been used to examine the language of the emotional speech recognition system.

Tools used for Evaluation

The main and most useful statistical measurements that were utilized to evaluate the emotional speech recognition on mobile phones are: recall, precision and accuracy. These measures are critical in that they are key elements in judging the performance of the emotional speech recognition. The recall rate measures and studies the relationship between the correct classification rates of specific emotions and the wrong classification of this specific emotion whereas precision measures the relation between the correct classification rate of specific emotions and the wrong classification of other emotions that are classified as special emotions. Finally, the
accuracy rate measures the relation between the correct classification rate of specific and other emotions and the total number of tested speech instances files.

Experimental Results

The performance of Mel-Frequency Cepstral Coefficient (MFCC) feature extracted and Support Vector Machine (SVM) algorithm has been examined. The experiment focuses on three speech emotional datasets, namely The Ryerson Audio-Visual Database of Emotional Speech and Song Dataset, Ryerson Multimedia Research Lab dataset, and Surrey Audio-Visual Expressed Emotion dataset. All vocal expressions in the three datasets were extracted with speech detector of speech recognition system in Android platform library. The system has been tested over the wave files (with 16bit and 48kHz). The previous source has been used separately. To avoid confusion, firstly, the emotion recognition system has been tested using wave files. In this testing, the system has examined its ability to distinguish between seven types of emotion: natural, happiness, sad, angry, fear, disgust and surprise. It is noticed that it has been done by using three different datasets. The following tables display the results that are collected from the three datasets files. Table 15, demonstrates the system results of (RAVDESS) dataset which contains 2880 files has been used for testing purposes. This dataset is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. It contains 8 emotions such as Calm, neutral, happy, sad, angry, fear, disgust and surprised emotions. Table 16, demonstrates the system results of (RML) dataset which contains 720 audio-visual emotional expression samples, and the video samples were collected from eight human subjects, speaking six different languages and six basic human emotions are expressed. Table 17, demonstrates the system results of (SAVEE) dataset contains 405 different British speeches from four male actors, aged between 27-31, in seven different emotions. The formal that been used to calculate the accuracy for these three tables is:
\[
\text{Accuracy} = \frac{\text{correct predicted}}{\text{total number of images}} \times 100\% \quad (36)
\]

Table 15: Confusion matrix SVM classification and number of wave files in the 1st dataset of Ryerson Audio-Visual Database of Emotional Speech and Song Dataset (RAVDESS).

<table>
<thead>
<tr>
<th>Predicted</th>
<th>True</th>
<th>N</th>
<th>F</th>
<th>Su</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>C</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral (N)</td>
<td></td>
<td>91.32</td>
<td>0</td>
<td>1.75</td>
<td>0</td>
<td>1.75</td>
<td>0</td>
<td>3.51</td>
<td>1.75</td>
<td></td>
</tr>
<tr>
<td>Fear (F)</td>
<td></td>
<td>0.87</td>
<td>96.52</td>
<td>1.74</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>Surprise(S)</td>
<td></td>
<td>0.87</td>
<td>0</td>
<td>95.65</td>
<td>0.87</td>
<td>0.87</td>
<td>0</td>
<td>0.87</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>Happy (H)</td>
<td></td>
<td>1.74</td>
<td>0</td>
<td>0</td>
<td>95.65</td>
<td>0.87</td>
<td>0.87</td>
<td>0</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>Sad (S)</td>
<td></td>
<td>1.74</td>
<td>0.87</td>
<td>0</td>
<td>96.52</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Angry (A)</td>
<td></td>
<td>0</td>
<td>0.87</td>
<td>0.87</td>
<td>0</td>
<td>0.87</td>
<td>95.65</td>
<td>0.87</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>Disgust (D)</td>
<td></td>
<td>0.87</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.87</td>
<td>97.39</td>
<td>0.87</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>Calm(C)</td>
<td></td>
<td>0.87</td>
<td>0</td>
<td>0.87</td>
<td>0.87</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>97.39</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>95.76%</td>
</tr>
</tbody>
</table>

Table 16: Confusion matrix SVM classification and number of wave files in the 2nd dataset of Ryerson Multimedia Research Lab (RML).

<table>
<thead>
<tr>
<th>Predicted</th>
<th>True</th>
<th>F</th>
<th>Su</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear (F)</td>
<td></td>
<td>91.67</td>
<td>2.78</td>
<td>2.78</td>
<td>0</td>
<td>2.78</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Surprise(S)</td>
<td></td>
<td>2.78</td>
<td>86.11</td>
<td>0</td>
<td>5.56</td>
<td>2.78</td>
<td>2.78</td>
<td></td>
</tr>
<tr>
<td>Happy (H)</td>
<td></td>
<td>2.78</td>
<td>0</td>
<td>88.89</td>
<td>2.78</td>
<td>2.78</td>
<td>2.78</td>
<td></td>
</tr>
<tr>
<td>Sad (S)</td>
<td></td>
<td>2.78</td>
<td>0</td>
<td>0</td>
<td>97.22</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Angry (A)</td>
<td></td>
<td>0</td>
<td>0</td>
<td>8.33</td>
<td>2.78</td>
<td>86.11</td>
<td>2.78</td>
<td></td>
</tr>
<tr>
<td>Disgust (D)</td>
<td></td>
<td>2.78</td>
<td>2.78</td>
<td>2.78</td>
<td>0</td>
<td>2.78</td>
<td>88.89</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>89.82%</td>
</tr>
</tbody>
</table>
Table 17: Confusion matrix SVM classification and number of wave files in the 3th dataset of Surrey Audio-Visual Expressed Emotion (SAVEE).

<table>
<thead>
<tr>
<th>Predicted</th>
<th>N</th>
<th>F</th>
<th>Su</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral (N)</td>
<td>91.67</td>
<td>2.78</td>
<td>2.78</td>
<td>0</td>
<td>2.78</td>
<td>0</td>
<td>0</td>
<td>79.76%</td>
</tr>
<tr>
<td>Fear (F)</td>
<td>0</td>
<td>83.33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11.11</td>
<td>5.56</td>
<td></td>
</tr>
<tr>
<td>Surprise(S)</td>
<td>0</td>
<td>0</td>
<td>83.33</td>
<td>5.56</td>
<td>11.11</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Happy (H)</td>
<td>0</td>
<td>5.56</td>
<td>0</td>
<td>77.78</td>
<td>0</td>
<td>11.11</td>
<td>5.56</td>
<td></td>
</tr>
<tr>
<td>Sad (S)</td>
<td>5.56</td>
<td>5.56</td>
<td>11.11</td>
<td>5.56</td>
<td>72.22</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Angry (A)</td>
<td>0</td>
<td>11.11</td>
<td>16.67</td>
<td>0</td>
<td>0</td>
<td>72.22</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Disgust (D)</td>
<td>0</td>
<td>0</td>
<td>5.56</td>
<td>5.56</td>
<td>11.11</td>
<td>0</td>
<td>77.78</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>0</td>
<td>0</td>
<td>5.56</td>
<td>5.56</td>
<td>11.11</td>
<td>0</td>
<td>77.78</td>
<td>79.76%</td>
</tr>
</tbody>
</table>

Note that the numbers of the test instances for each emotion are not equal because the instances for testing have been selected randomly from the test instances. Moreover, the system may test an instance more than once.

Moving to a more accurate result, the tables below illustrate the result of the system over the database file that contains whole image training and testing. The method that been used to calculate these three tables is: Recall, Precision, and Accuracy. The system behavior over three datasets that have been examined and using 70% as training wave files with 30% as testing wave files for each dataset. The first dataset has 2880 instances, the second also has 720 instances, and the third dataset has 405 instances. The following tables demonstrate the system performance with each group of data.
Table 18: Confusion matrix using SVM classification for RAVDESS dataset.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>F</th>
<th>Su</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>N</th>
<th>C</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear</td>
<td>111</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>98.23</td>
<td>96.52</td>
<td>99.17</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
<td>0</td>
<td>110</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>96.49</td>
<td>95.65</td>
<td>98.76</td>
</tr>
<tr>
<td>Happy</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>111</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>97.35</td>
<td>95.65</td>
<td>98.89</td>
</tr>
<tr>
<td>Sad</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>110</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>98.21</td>
<td>95.65</td>
<td>99.03</td>
</tr>
<tr>
<td>Angry</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>112</td>
<td>0</td>
<td>1</td>
<td>96.55</td>
<td>97.39</td>
<td>99.37</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>52</td>
<td>1</td>
<td>88.14</td>
<td>91.23</td>
<td>98.35</td>
</tr>
<tr>
<td>Neutral</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>112</td>
<td>1</td>
<td>94.92</td>
<td>97.39</td>
<td>98.75</td>
</tr>
<tr>
<td>Calm</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>112</td>
<td>95.59%</td>
<td>95.75%</td>
<td>98.83%</td>
</tr>
</tbody>
</table>

Table 19: Confusion matrix using SVM classification for RML dataset.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>F</th>
<th>Su</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear</td>
<td>33</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>89.18</td>
<td>91.6</td>
<td>96.52</td>
</tr>
<tr>
<td>Surprise</td>
<td>1</td>
<td>31</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>93.93</td>
<td>86.1</td>
<td>96.51</td>
</tr>
<tr>
<td>Happy</td>
<td>1</td>
<td>0</td>
<td>32</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>86.48</td>
<td>88.8</td>
<td>95.57</td>
</tr>
<tr>
<td>Sad</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>35</td>
<td>0</td>
<td>0</td>
<td>89.74</td>
<td>97.2</td>
<td>97.48</td>
</tr>
<tr>
<td>Angry</td>
<td>0</td>
<td>0</td>
<td>31</td>
<td>1</td>
<td>31</td>
<td>1</td>
<td>88.57</td>
<td>86.1</td>
<td>95.57</td>
</tr>
<tr>
<td>Disgust</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>32</td>
<td>1</td>
<td>91.43</td>
<td>88.8</td>
<td>96.52</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>89.88%</td>
<td>89.76%</td>
<td>96.36%</td>
</tr>
</tbody>
</table>
Table 20: Confusion matrix using SVM classification for SAVVE dataset.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>F</th>
<th>S</th>
<th>H</th>
<th>A</th>
<th>D</th>
<th>N</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>75</td>
<td>83.3</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
<td>15</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>68.2</td>
<td>83.3</td>
</tr>
<tr>
<td>Happy</td>
<td>1</td>
<td>0</td>
<td>14</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>82.35</td>
<td>77.7</td>
</tr>
<tr>
<td>Sad</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>72.2</td>
<td>72.2</td>
</tr>
<tr>
<td>Angry</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>76.47</td>
<td>72.2</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>87.5</td>
<td>77.7</td>
</tr>
<tr>
<td>Neutral</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>97.05</td>
<td>91.6</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>79.82%</td>
<td>79.71%</td>
</tr>
</tbody>
</table>

Table 21: Total number of wave files collected for speech emotion study.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Total Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name of Dataset</td>
<td>RAVDESS</td>
</tr>
<tr>
<td></td>
<td>Train -70%</td>
</tr>
<tr>
<td>Happy</td>
<td>266</td>
</tr>
<tr>
<td>Sad</td>
<td>263</td>
</tr>
<tr>
<td>Fear</td>
<td>267</td>
</tr>
<tr>
<td>Surprise</td>
<td>265</td>
</tr>
<tr>
<td>Anger</td>
<td>267</td>
</tr>
<tr>
<td>Disgust</td>
<td>265</td>
</tr>
<tr>
<td>Neutral</td>
<td>128</td>
</tr>
<tr>
<td>Calm</td>
<td>263</td>
</tr>
<tr>
<td>Total</td>
<td>2880 wave files</td>
</tr>
</tbody>
</table>

5.3. Choosing Base Model for Information Fusion

In the Information Fusion approach, we have already built many different models based on different databases, feature sets, as well as the classification algorithm as we did same for speech emotion recognition. According to the conclusion, the RAVDESS, RML and SAVVE datasets based on Support Vector Machine (SVM)
algorithm. The following tests will also show how the ethnicity and language affect the accuracy of the facial expression recognition and the speech emotional recognition.

Usage Dataset for Evaluation

A set of 2880 instances has been used for testing purposes. This dataset is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. It contains eight emotions such as Calm, neutral, happy, sad, angry, fear, disgust and surprised emotions. In this experiment, the RAVDESS database has been used to examine the language and ethnicity of the facial expression recognition and the speech emotional recognition. Three types of information fusion are tested in this section. Table 22, 23 and 24 show the confusion matrices of the information fusion. The recognition rates are 97.38%, 94.90% and 90.38%, respectively.

Table 22: Confusion matrix using SVM classification for RAVDESS dataset.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>F</th>
<th>Su</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>C</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral (N)</td>
<td>96.49</td>
<td>1.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.75</td>
<td>97.38%</td>
</tr>
<tr>
<td>Fear (F)</td>
<td>0.87</td>
<td>97.39</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.87</td>
<td>0.87</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Surprise(S)</td>
<td>0.87</td>
<td>0</td>
<td>97.39</td>
<td>0</td>
<td>0.87</td>
<td>0</td>
<td>0</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>Happy (H)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>97.39</td>
<td>0.87</td>
<td>0</td>
<td>0.87</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sad (S)</td>
<td>0</td>
<td>0</td>
<td>1.74</td>
<td>0</td>
<td>97.39</td>
<td>0.87</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Angry (A)</td>
<td>0</td>
<td>0.87</td>
<td>0</td>
<td>0</td>
<td>0.87</td>
<td>97.39</td>
<td>0.87</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Disgust (D)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.87</td>
<td>0</td>
<td>0</td>
<td>97.39</td>
<td>1.74</td>
<td></td>
</tr>
<tr>
<td>Calm(C)</td>
<td>0</td>
<td>0</td>
<td>0.87</td>
<td>0</td>
<td>0.87</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>98.26</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>97.38%</td>
</tr>
</tbody>
</table>
Table 23: Confusion matrix using SVM classification for RML dataset.

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>Su</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear (F)</td>
<td>94.44</td>
<td>2.78</td>
<td>2.78</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Surprise(S)</td>
<td>0</td>
<td>94.44</td>
<td>0</td>
<td>0</td>
<td>5.56</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Happy (H)</td>
<td>2.78</td>
<td>0</td>
<td>94.44</td>
<td>0</td>
<td>2.78</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sad (S)</td>
<td>0</td>
<td>0</td>
<td>2.78</td>
<td>94.44</td>
<td>0</td>
<td>2.78</td>
<td></td>
</tr>
<tr>
<td>Angry (A)</td>
<td>0</td>
<td>0</td>
<td>2.78</td>
<td>2.78</td>
<td>94.44</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Disgust (D)</td>
<td>0</td>
<td>0</td>
<td>2.78</td>
<td>0</td>
<td>0</td>
<td>97.22</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>94.90%</td>
</tr>
</tbody>
</table>

Table 24: Confusion matrix using SVM classification for SAVEE dataset.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>F</th>
<th>Su</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral (N)</td>
<td>92.59</td>
<td>0</td>
<td>3.70</td>
<td>0</td>
<td>3.70</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Fear (F)</td>
<td>0</td>
<td>87.50</td>
<td>6.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.25</td>
<td></td>
</tr>
<tr>
<td>Surprise(S)</td>
<td>0</td>
<td>0</td>
<td>92.31</td>
<td>0</td>
<td>7.69</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Happy (H)</td>
<td>0</td>
<td>0</td>
<td>13.33</td>
<td>86.76</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sad (S)</td>
<td>0</td>
<td>0</td>
<td>15.38</td>
<td>84.62</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Angry (A)</td>
<td>5.56</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>94.44</td>
<td></td>
</tr>
<tr>
<td>Disgust (D)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.56</td>
<td>0</td>
<td>94.44</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>90.38%</td>
</tr>
</tbody>
</table>

Note that the numbers of the test instances for each emotion are not equal because the instances for testing have been selected randomly from the test instances. Moreover, the system may test a file more than once.

Moving to a more accurate result, the tables below illustrate the result of the system over the database file that contains whole image training and testing. The method that been used to calculate these three tables is: Recall, Precision, and Accuracy. The system behavior over three datasets that have been examined and using 70% as training images with 30% as testing images for each dataset. The first dataset has 2880 instances, the second also has 720 instances, and the third dataset has 405
instances. The following tables demonstrate the system performance with each group of data.

Table 25: Confusion matrix using SVM classification for RAVDESS dataset.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>F</th>
<th>Su</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>N</th>
<th>C</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear</td>
<td>112</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>98.25</td>
<td>98.25</td>
<td>99.52</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
<td>112</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>97.39</td>
<td>97.39</td>
<td>99.29</td>
</tr>
<tr>
<td>Happy</td>
<td>0</td>
<td>0</td>
<td>112</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>99.12</td>
<td>97.39</td>
<td>99.53</td>
</tr>
<tr>
<td>Sad</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>112</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>96.55</td>
<td>97.39</td>
<td>99.17</td>
</tr>
<tr>
<td>Angry</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>112</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>98.25</td>
<td>96.55</td>
<td>99.41</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>112</td>
<td>1</td>
<td>2</td>
<td>97.39</td>
<td>96.55</td>
<td>99.17</td>
</tr>
<tr>
<td>Neutral</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>55</td>
<td>1</td>
<td>96.49</td>
<td>96.49</td>
<td>99.53</td>
</tr>
<tr>
<td>Calm</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>113</td>
<td>95.76</td>
<td>98.26</td>
<td>99.17</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>97.4%</strong></td>
<td><strong>97.28%</strong></td>
<td><strong>99.35%</strong></td>
</tr>
</tbody>
</table>

Table 26: Confusion matrix using SVM classification for RML dataset.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>F</th>
<th>Su</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear</td>
<td>34</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>97.14</td>
<td>94.4</td>
<td>98.55</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
<td>34</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>97.14</td>
<td>94.4</td>
<td>98.55</td>
</tr>
<tr>
<td>Happy</td>
<td>1</td>
<td>0</td>
<td>34</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>89.47</td>
<td>94.4</td>
<td>97.14</td>
</tr>
<tr>
<td>Sad</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>34</td>
<td>0</td>
<td>1</td>
<td>97.14</td>
<td>94.4</td>
<td>98.55</td>
</tr>
<tr>
<td>Angry</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>34</td>
<td>0</td>
<td>91.89</td>
<td>94.4</td>
<td>97.61</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>34</td>
<td>97.14</td>
<td>97.1</td>
<td>99.02</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>94.9%</strong></td>
<td><strong>94.85%</strong></td>
<td><strong>98.24%</strong></td>
</tr>
</tbody>
</table>
Table 27: Confusion matrix using SVM classification for SAVEE dataset.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>F</th>
<th>Su</th>
<th>H</th>
<th>S</th>
<th>A</th>
<th>D</th>
<th>N</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear</td>
<td>14</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>87.5</td>
<td>98.19</td>
</tr>
<tr>
<td>Surprise</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>66.6</td>
<td>92.3</td>
<td>93.96</td>
</tr>
<tr>
<td>Happy</td>
<td>0</td>
<td>2</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>86.6</td>
<td>98.19</td>
</tr>
<tr>
<td>Sad</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>84.62</td>
<td>84.6</td>
<td>96.46</td>
</tr>
<tr>
<td>Angry</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>94.4</td>
<td>99.09</td>
</tr>
<tr>
<td>Disgust</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>94.4</td>
<td>94.4</td>
<td>98.19</td>
</tr>
<tr>
<td>Neutral</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>92.59</td>
<td>92.6</td>
<td>96.46</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>91.17%</strong></td>
<td><strong>90.34%</strong></td>
<td><strong>97.22%</strong></td>
</tr>
</tbody>
</table>

Table 28: Total number of wave files collected for each part of the study.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Facial Expression</th>
<th>Speech Emotion</th>
<th>Information Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CK+ Train 70%</td>
<td>RAVDESS Train 70%</td>
<td>RAVDESS Test 30%</td>
</tr>
<tr>
<td>Happy</td>
<td>49 train 70%</td>
<td>266 train 70%</td>
<td>266 test 30%</td>
</tr>
<tr>
<td>Sad</td>
<td>20 train 70%</td>
<td>263 train 70%</td>
<td>263 test 30%</td>
</tr>
<tr>
<td>Fear</td>
<td>18 train 70%</td>
<td>267 train 70%</td>
<td>267 test 30%</td>
</tr>
<tr>
<td>Surprise</td>
<td>59 train 70%</td>
<td>265 train 70%</td>
<td>265 test 30%</td>
</tr>
<tr>
<td>Anger</td>
<td>32 train 70%</td>
<td>267 train 70%</td>
<td>267 test 30%</td>
</tr>
<tr>
<td>Disgust</td>
<td>42 train 70%</td>
<td>265 train 70%</td>
<td>265 test 30%</td>
</tr>
<tr>
<td>Neutral</td>
<td>229 train 70%</td>
<td>128 train 70%</td>
<td>128 test 30%</td>
</tr>
<tr>
<td>Calm</td>
<td>-</td>
<td>263 train 70%</td>
<td>263 test 30%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>636 images</strong></td>
<td><strong>2880 files</strong></td>
<td><strong>2880 images/files</strong></td>
</tr>
</tbody>
</table>
Table 29: The detailed accuracy of emotion recognition.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Facial expression</th>
<th>Speech Information</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>100%</td>
<td>95.65%</td>
<td>97.39%</td>
</tr>
<tr>
<td>Sad</td>
<td>87.5%</td>
<td>95.52%</td>
<td>97.39%</td>
</tr>
<tr>
<td>Fear</td>
<td>100%</td>
<td>96.52%</td>
<td>97.39%</td>
</tr>
<tr>
<td>Surprise</td>
<td>100%</td>
<td>95.65%</td>
<td>97.39%</td>
</tr>
<tr>
<td>Anger</td>
<td>92.3%</td>
<td>95.65%</td>
<td>97.39%</td>
</tr>
<tr>
<td>Disgust</td>
<td>94.1%</td>
<td>97.39%</td>
<td>97.39%</td>
</tr>
<tr>
<td>Neutral</td>
<td>100%</td>
<td>91.32%</td>
<td>96.49%</td>
</tr>
</tbody>
</table>

| Overall Rec Rate | 96.3% | 95.43% | 97.26% |

Through information fusion, the improvement of emotion recognition can achieve 1.83% compared to the speech emotion recognition and 0.96% compared to the facial emotion recognition. Table 29 shows the detail comparison of different emotion recognition. From the figure we can see that the information fusion has better results for most of the emotions. In addition, the overall performance is better than facial expression and speech information. Table 29 is the detail information of Figure 56.
Figure 56: The comparison of multi-sensory emotion recognition.
5.4. The Experiment Results Discussion

The final step in this research is to assess and evaluate the project performance; to measure how many of the requirements for emotional face and speech recognition on mobile phones using cloud computing can be achieved. Actually, testing has been continuously addressed from the early implementation stage until the final stage.

Firstly, the testing of each function is carried out individually. It is tested to ensure that the algorithm and each line code works correctly. Sometimes, a small sample of the image/wave files and both data are used to test the code and sometimes a different kind of data set is built. For example, a database of numeric matrix is built to calculate easily the results of the tested function manually and to compare the target results with the code results.

Secondly, after completing a certain stage, the performance of that stage is tested. Furthermore, after integrating the system stages, the overall system performance was tested. In these phases, sometimes an implemented algorithm that is not useful for this project is discovered; then one is forced to return to the starting point. For example, the support vector machines (SVM) algorithm is described three times in three different papers. All these attempts failed in our hands. The problem was with the huge number of multi-class features that need to be trained. To solve this problem, attention was turned to the Android platform tools and a MATLAB tool that can be used with the project data. The Android platform was used to program the application and the MATLAB tool used to simulate the application.

Finally, the entire project was tested to guarantee that the proposed method met the requirements and achieved the aim. This is the most significant step, which ensures that the study algorithm is efficient and powerful; that it could recognize emotional speech and face expressions from a mobile phone using cloud computing that run on real time and you can use it anywhere you go.
Table 30 displays the correlative properties of our proposed study and the previous research in facial expression recognition. We applied the methods we put forth in our research as well as the previous research methods to the three databases of CK+, JAFFE, and KDEF. We found that there are some definitive factors that affect the different behaviors between our proposed methods and the earlier, documented methods of research. These factors encompass things such as previously, accuracy of outcome was determined by examining the highest number of performed expressions, whereas in our proposed method, we determined comprehensive accuracy by juxtaposing the total image number beside the amount of predictions that were correct. In our study, we employed all of the database images available to us, while in previous research, use of every database image was not accomplished.

Our suggested method of application of facial and speech emotion in a cloud environment on a mobile phone was successful. However, by implementing different features in the app, our work can be taken beyond this study. Intelligent expansion of our research would include features such as emotion recognition via speech; if the system can perform a combination of voice/speech recognition along with facial expression recognition, then this information could possibly be used in various security applications, such as criminal exposure and apprehension, and airport security measures.

### Table 30: Accuracy Comparison Between the Systems.

<table>
<thead>
<tr>
<th>Systems</th>
<th>CK+</th>
<th>JAFFE</th>
<th>KDEF</th>
</tr>
</thead>
<tbody>
<tr>
<td>[71]</td>
<td>83.2</td>
<td>-</td>
<td>74.6</td>
</tr>
<tr>
<td>[72]</td>
<td>89.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[73]</td>
<td>81.4</td>
<td>-</td>
<td>82.4</td>
</tr>
<tr>
<td>[74]</td>
<td>85.4</td>
<td>96.4</td>
<td>-</td>
</tr>
<tr>
<td>[75]</td>
<td>81.6</td>
<td>94.9</td>
<td>-</td>
</tr>
<tr>
<td>[76]</td>
<td>95.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Proposed</td>
<td>96.3</td>
<td>91.9</td>
<td>90.8</td>
</tr>
</tbody>
</table>
CHAPTER 6: CONCLUSION AND FUTURE WORK

This dissertation has established a facial expression and speech emotion recognition system that been developed for smartphone by integrating multi-channel information fusion. The goal of the proposed system is to investigating and improve the emotion recognition accuracy on mobile phone using cloud computing.

Conclusion

In this dissertation, the information fusion framework was proposed which mainly contains three parts: the speech emotion recognizer, the facial emotion recognizer, and information fusion. The proposed approach tries to distinguish between six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) as well as the neutral and calm state by integrating information from audio (speech) and video (facial expression) channels. The proposed framework can also do either speech emotion recognition or facial expression recognition individually.

Emo application is used in speech emotion recognition to process the audio file, extract acoustic features, and classify the emotions. Three emotional databases RAVDESS, SAVEE and RML were used to find the base model. We have tested three types of features 384 acoustic features, 120 acoustic features and 45 acoustic features. Experiments demonstrate that models with 384 features have better performance than 120 features and 45 features. one type of classification algorithm was used which is SVM algorithm. The recognition accuracy of this chosen base model can achieve 95.76%, 89.82% and 79.76% for RAVDESS, RML and SAVEE, respectively. The experiment of the speech emotion recognition also indicates that emotion is language dependent. It is hard to have a general model for speech emotion recognition with a different language.

Two approaches were achieved for the facial emotion recognition. One is used to recognize facial emotion from static images. In this approach, skin color
segmentation was used for the face detection was used for the feature extraction. The CK+, JAFFE, and KDEF databases were used to test the accuracy of this approach. The recognition accuracy of this chosen base model can achieve 96.3%, 93.03% and 90.8% for CK+, JAFFE, and KDEF respectively when using 80% for the training purposes and 20% for testing. The experiments demonstrate that the more the training data, the better the accuracy it can achieve. Another approach is recognizing emotion from live video on smartphone. The ASM tracker was used for 68 facial landmarks. Seven high-level regions were extracted from the 68 facial landmarks as the feature set. The nearest distance is applied as the classification method. Though the training model is better if it is person specific, a generic model was built for interacting with different subjects. For the best use of the facial emotion recognizer, one may first train a person-specific model and then use it to recognize emotions.

The instances of the RAVDESS database are used to test the multi-sensory emotion recognition framework. In the speech emotion recognition tests, the recognition accuracy can achieve 95.43% on smartphone. The average facial expression recognition rate is 96.3%. There are one types of method used in the information fusion. One fuses the speech and facial emotion recognition results directly. The improvement of the recognition is 1% compared to the facial expression recognition and 1.83% compared to the speech emotion recognition.

One thing to be pointed out is that, though the result of the facial emotion recognition is slightly higher than the speech, it can achieve even higher result if the test instance is much closer to our requirement. Our requirement for the facial video is that at the beginning, the testing subject should have a neutral state for the initialization. However, most of the test instances do not satisfy this requirement, and we can only use the mean shape of the first frame (is the initialization threshold) of the test subject to initialize. Thus, the new audio-visual emotional database is needed to demonstrate the advantage of the proposed system in the future.
Future Work

Based on our research results in this dissertation, more opportunities for future research in this field can be done to extend this work. First, more emotional data samples can be collected from a large number of people, and these samples can be used in the speech and facial emotion recognition for the training purpose or testing purpose and also to form an emotional database to convenient other researchers in this area. If such a database can be collected, it may become a milestone in the field of emotion recognition.

Second, it is possible to add feedback to the facial and speech emotion recognition framework to improve the recognition accuracy by assigning a dynamic weight to the information fusion instead of using the fixed weight. Another possible extension is to integrate more modalities into the system such as body gesture, hand gesture, etc. Adding more modalities can provide more information about the emotional states of human being, thus, may achieve better emotion recognition accuracy. Finally, different acoustic features can be tested to improve the robustness of the base model in speech emotion recognition.

The facial and speech emotion recognition framework proposed in this dissertation can be extended to many other research topics in the field of human-computer interaction. We hope our research can trigger more investigations in the area of human-computer interaction to make computers and smartphones more friendly, natural, and human-like shortly.
References


[110] Ying-Li Tian Takeo Kanade and Jeffrey F. Cohn, Facial Expression Analysis.


[112] Ying-Li Tian Takeo Kanade and Jeffrey F. Cohn, Facial Expression Analysis.


156


Appendix A: Publications

Conference Papers


Journal Papers


