






<https://dx.doi.org/10.17488/RMIB.45.2.1>

E-LOCATION ID: 1408

# Gammatone-Frequency Cepstral Coefficients Based Fear Emotion Level Recognition System

## Sistema de Reconocimiento de Nivel de Emoción Basado en Coeficientes Cepstrales de Frecuencia Gammatone

Barlian Henryranu Prasetio<sup>1</sup>  , La Ode Adriyan Hazmar<sup>2</sup> , Dahnia Syauqy<sup>1</sup> , Edita Rosana Widasari<sup>1</sup> 

<sup>1</sup>Universitas Brawijaya, Faculty of Computer Science - Indonesia

<sup>2</sup>Universitas Brawijaya, Computer Engineering- Indonesia

### ABSTRACT

Emotions represent affective states that induce alterations in behavior and interactions within one's environment. An avenue for discerning human emotions lies in the realm of speech analysis. Empirical evidence indicates that 1.6 million Indonesian teenagers grapple with mental anxiety disorders, characterized by sensations of fear or ambiguous vigilance. This work endeavors to devise a tool for discerning an individual's emotional state through voice processing, focusing particularly on fear emotions stratified into three levels of intensity: low, medium, and high. The proposed system employs Gammatone-Frequency Cepstral Coefficients (GFCC) for feature extraction, leveraging the efficacy of its gamma filter in reducing noise. Furthermore, a Random Forest (RF) Classifier is integrated to facilitate the recognition of fear's emotional intensity in speech signals. The system is deployed on a Raspberry Pi 4B and establishes a Bluetooth connection using the RFCOMM communication protocol to an Android application, presenting the classification results. The outcomes reveal that the Signal-to-Noise Reduction achieved through GFCC extraction surpasses that of Mel-Frequency Cepstral Coefficients (MFCC). In terms of accuracy, the implemented recognition system for fear emotion levels, employing GFCC extraction and Random Forest Classifier, attains a commendable accuracy of 73.33 %.

**KEYWORDS:** fear emotion, gammatone-frequency cepstral coefficients, Mel-frequency cepstral coefficients, signal-to-noise reduction, speech sound

## RESUMEN

Las emociones representan estados afectivos que inducen alteraciones en el comportamiento e interacciones dentro del entorno de un individuo. Un enfoque para discernir las emociones humanas se encuentra en el análisis del habla. La evidencia empírica indica que 1.6 millones de adolescentes indonesios enfrentan trastornos de ansiedad mental, caracterizados por sensaciones de miedo o vigilancia ambigua. Esta investigación se propone diseñar una herramienta para discernir el estado emocional de una persona mediante el procesamiento de la voz, centrándose especialmente en las emociones de miedo estratificadas en tres niveles de intensidad: bajo, medio y alto. La metodología propuesta emplea los Coeficientes Cepstrales de Frecuencia Gammatone (GFCC) para la extracción de características, aprovechando la eficacia de su filtro gamma para combatir el ruido. Además, se incorpora un Clasificador Random Forest (RF) para facilitar el reconocimiento de la intensidad emocional del miedo en las señales de voz. El sistema se implementa en una Raspberry Pi 4B y establece una conexión Bluetooth utilizando el protocolo de comunicación RFCOMM con una aplicación Android, presentando los resultados de la clasificación. Los resultados revelan que la Reducción de Señal a Ruido lograda mediante la extracción de GFCC supera a la de los Coeficientes Cepstrales de Frecuencia Mel (MFCC). En términos de precisión, el sistema de reconocimiento implementado para los niveles de emoción de miedo, utilizando la extracción de GFCC y el Clasificador Random Forest, alcanza una precisión destacada del 73.33 %

**PALABRAS CLAVE:** emoción de miedo, coeficientes cepstrales de frecuencia gammatone, coeficientes cepstrales de frecuencia Mel, reducción de señal a ruido, sonido del habla

### Corresponding author

TO: Barlian Henryranu Prasetio

INSTITUTION: Universitas Brawijaya, Faculty of Computer  
Science - Indonesia

ADDRESS: Jl. Veteran, Ketawanggede, Kec. Lowokwaru,  
Kota Malang, Jawa Timur 65113, Indonesia

CORREO ELECTRÓNICO: barlian@ub.ac.id

### Received:

27 November 2023

### Accepted:

3 April 2024

## INTRODUCTION

Speech serves as a communicative medium for conveying information in a manner comprehensible to others <sup>[1]</sup>. It is shaped by a signal influenced by both time and frequency. Speech possesses distinct characteristics, including pitch, voice type, timbre, and volume <sup>[2]</sup>. Speech encompasses two content components: verbal and nonverbal. Verbal content comprises words interpreted by listeners, while nonverbal content encapsulates information conveyed through the way these words are expressed. Within the speech, nonverbal content holds the potential to communicate an individual's emotional state <sup>[3]</sup>.

Nowadays, being aware of our emotions is crucial for personal well-being and effective navigation through life's complexities <sup>[4]</sup>, a principle that holds true in Indonesia. It serves as the cornerstone of self-understanding, enabling individuals to discern the intricacies of their feelings, identify patterns, and foster personal growth. Emotional awareness enhances communication skills, allowing for clear expression of thoughts and feelings, minimizing the likelihood of misunderstandings in relationships <sup>[5]</sup>. In conflicts, this awareness facilitates empathetic resolution and constructive problem-solving. Moreover, it plays a pivotal role in stress management, as recognizing emotional triggers empowers individuals to implement coping strategies and maintain mental health. By embracing emotional awareness, we bolster decision-making processes, navigate challenges with resilience, and cultivate empathy and compassion for others. Ultimately, fostering emotional awareness is integral to creating a positive emotional climate, promoting overall well-being, and building meaningful connections our self to others.

Emotion awareness, as illuminated by Robert Plutchik's emotional wheel, holds profound significance in the realm of stress management, especially considering survey results revealing that 1.6 million Indonesian teenagers grapple with mental anxiety disorders marked by feelings of fear or ambiguous vigilance <sup>[6]</sup>. Emotion, as an intrinsic human experience and a reciprocal reaction to actions, situations, or events, frequently gives rise to behavioral and interactive changes within the surrounding environment. Plutchik's model, categorizing basic emotions into eight distinct parts and further delineating them into three intensity levels: low, medium, and high <sup>[7]</sup>, serves as a valuable tool for comprehending the intricate emotional landscape associated with anxiety disorders. Integrating these perspectives emphasizes the interconnected nature of emotions, stress, and mental health, highlighting the importance of emotion awareness in crafting tailored stress management strategies that acknowledge the nuanced levels of fear and vigilance experienced by individuals in specific contexts.

The correlation between high levels of fear emotion and stress is a well-established aspect of psychological and physiological responses to challenging situations. Fear, as a primal and adaptive emotion, triggers the body's "fight or flight" response, releasing stress hormones such as cortisol and adrenaline. In situations where fear is elevated, the body perceives a potential threat, leading to heightened physiological arousal and increased stress levels <sup>[8]</sup>. Chronic or intense fear can contribute to sustained stress, negatively impacting both mental and physical well-being. Recognizing and understanding this connection between fear and stress is crucial in developing effective stress management strategies. Emotion awareness, particularly concerning fear at varying intensity levels, becomes instrumental in tailoring interventions and coping mechanisms to address the specific emotional challenges contributing to elevated stress levels in individuals.

In general, apart from being expressed verbally, emotions are intricately tied to physiological. Physical responses, including changes in muscle tension, especially head and neck area. For instance, heightened emotions, such as

stress or fear, may result in increased tension in the neck muscles, significantly impacting speech characteristics. Speech stands out as a preeminent method for recognizing and comprehending human emotions due to its rich and diverse set of communicative elements. The nuances in pitch, rhythm, intonation, and various vocal cues embedded in speech offer profound insights into an individual's emotional state [9]. This comprehensive spectrum of prosodic features, coupled with the tone, timbre, and non-verbal vocal cues, provides a multifaceted tapestry of emotional expression.

Emotions can undergo sudden changes in response to various circumstances, posing significant challenges when measuring them, especially unconscious emotions such as fear. This unpredictability can lead to discomfort in individuals, resulting in inaccurate measurements. Hence, non-invasive methods, particularly speech analysis [10], play a crucial role in accurately assessing emotional levels. non-invasive approach allows for emotion measurement without the need for physical contact or potentially disruptive procedures. Not only does this non-intrusive method uphold individual privacy, but it also establishes a more comfortable environment, enhancing the accuracy of emotion measurement, particularly in the face of abrupt emotional shifts.

Therefore, in this work, we propose to develop a system for recognizing the level of fear emotion condition through speech analysis. The proposed system identifies the emotion of fear in three intensity levels: Apprehension (low), Fear (medium), and Terror (high). Previous research on a similar topic utilized the Mel-Frequency Cepstral Coefficients (MFCC) extraction method [11]. While MFCC proves efficient in quiet surroundings, its adaptability to noisy environments is limited. Consequently, we propose to employ the Gammatone Frequency Cepstral Coefficients extraction method, featuring an effective gamma filter tailored for sounds with high noise levels [12]. Additionally, the study's findings can offer insights into the performance of the Random Forest Classifier in recognizing the emotional intensity of fear in sound signals.

### Related research

In 2020, Wang proposed the application of Gammatone Frequency Cepstral Coefficients (GFCC) for Forensic Automatic Speaker Recognition (FASR) in comparison under noisy conditions. The system uses GFCC and integrated with Principal Component Analysis (PCA) algorithm applied on mandarin voice datasets with different levels of white noise. Based on the result, it showed that overall system based on GFCC has improvement over baseline Mel Frequency Cepstral Coefficients (MFCC) on the same conditions [13].

Previously, similar research related to Speech Emotion Recognition (SER) using GFCC has been proposed by Bharti, *et al.* in 2020 [14]. The authors introduced the use of GFCC, ALO (Ant Lion Optimization), and Multi-Class Support Vector Machine (MSVM) classification in developing Speech Emotion Recognition (SER) system. They used RAVDESS Ryerson Audiovisual data set of expression voice and song dataset that contains 7356 records, comprises 24 speech samples, and happy-sad-angry emotions. The experimental system was simulated and executed using MATLAB with GUI tool. The evaluation parameters consisted of AUC, MSE, SNR, FAR and FRR to be compared with MFCC and SVM. The result showed that GFCC+ALO+MSVM achieved 97 % accuracy, while MFCC+SVM only achieved 79.48 % accuracy.

Patni, *et al.* proposed Speech Emotion Recognition using several features including MFCC, GFCC, Chromagram and RMSE features. Those 42 extracted features (16 MFCC, 12 GFCC, 13 Chromagram, 1 RMSE) from Ryerson Audio-

Visual Database of Emotional Speech and Song (RAVDESS) were then classified using 2D-CNN which achieved overall accuracy of more than 92 % <sup>[15]</sup>.

The research proposed by Choudhary in 2021 used Gammatone cepstral coefficients for automatic speaker verification. To validate the system, they used different voices from different speakers then the system decided whether the voice comes from the same or different person. The result showed that the system can effectively recognize the speaker with high accuracy. In addition, they found out that by using Gammatone cepstral coefficients, speaker verification achieved significant improvement compared to other methods based on cepstral coefficients <sup>[16]</sup>.

Zheng proposed another speech emotion recognition system based on the combination of convolutional neural network and random forest in 2018. CNN was used to extract speech emotion features from spectrogram and RF was then used to classify the emotion. The result showed the combination of using random forest boost the performance result compared to only traditional CNN model <sup>[17]</sup>.

Another research by Hamsa in 2020 proposed an approach and framework for emotion recognition based on voice in noisy conditions. They used Wavelet Packet Transform (WPT) based on cochlear filter bank instead of gammatone filter bank and short-time Fourier transform. The features were then classified using Random Forest (RF) algorithm. The result showed the performance on three speech corpora in two languages in noisy conditions was better than other algorithms <sup>[18]</sup>.

## Fear Emotion

Emotions are categorized into basic emotions and advanced emotions. Basic emotions include joy, resignation, surprise, sadness, disgust, anger, anticipation, and fear.

Emotions are interconnected in various ways, influencing each other, and contributing to the overall emotional landscape. For example, fear is intricately connected to the experience of stress, forming a complex relationship between psychological and physiological responses. Fear, as an adaptive and primal emotion, triggers the body's stress response commonly known as the "fight or flight" reaction. When confronted with a perceived threat or danger, the body releases stress hormones <sup>[19]</sup>, including cortisol and adrenaline, preparing the individual to respond to the imminent challenge. In situations where fear is prolonged or intense, it can contribute significantly to chronic stress. The persistent activation of the stress response system can lead to various physiological and psychological consequences, impacting overall well-being. Physiologically, prolonged stress linked to fear can result in heightened blood pressure, increased heart rate, and tension in muscles. Psychologically, the continuous experience of fear-related stress can contribute to anxiety disorders, sleep disturbances, and other mental health challenges. Moreover, individuals may adopt coping mechanisms that, while initially helpful in managing fear, can contribute to ongoing stress if not addressed effectively.

The intensity level of fear is a crucial dimension that enhances our comprehension of this intricate emotion. Fear, a complex emotional state, spans a spectrum of intensities, encompassing mild apprehension, moderate fear, and intense terror. At the lower end of the spectrum, individuals may experience a subtle sense of unease or concern, reflecting cautious responses to potential threats without feeling overwhelmed. Moving to the middle intensity level, fear becomes more pronounced, marked by heightened alertness and a palpable sense of anxiety. Individuals

at this stage exhibit more discernible physical and emotional reactions indicative of a moderate fear state. At the highest intensity level, fear transforms into terror, an overwhelming and distressing emotional experience. Here, individuals grapple with an acute sense of imminent danger, triggering intense physiological responses like increased heart rate and a heightened fight-or-flight reaction. This nuanced understanding of fear's intensity levels is invaluable across disciplines such as psychology, neuroscience, and emotion recognition research, facilitating a more precise analysis and targeted approach to managing fear across varying degrees of intensity.

Previous study entitled "Emotional Processing of Fear: Exposure to Corrective Information" conducted by Edna B. Foa and Michael J. Kozak in 1986 presents the results of research on how individuals process the emotion of fear and how they manage that fear [20]. According to the results of this study, fear is an emotion that arises when individuals feel threatened or feel insecure.

### Gammatone-Frequency Cepstral Coefficients (GFCC)

Gammatone Frequency Cepstral Coefficients (GFCC) were used to perform feature extraction in this study. This method is a development of the Mel-frequency Cepstral Coefficients (MFCC). Both methods start with preprocessing and FFT, but GFCC uses a gammatone filter bank which then compresses the results using the cubic root operation. After that, it is processed using DCT [8][16]. Figure 1 depicts the flow of GFCC steps.

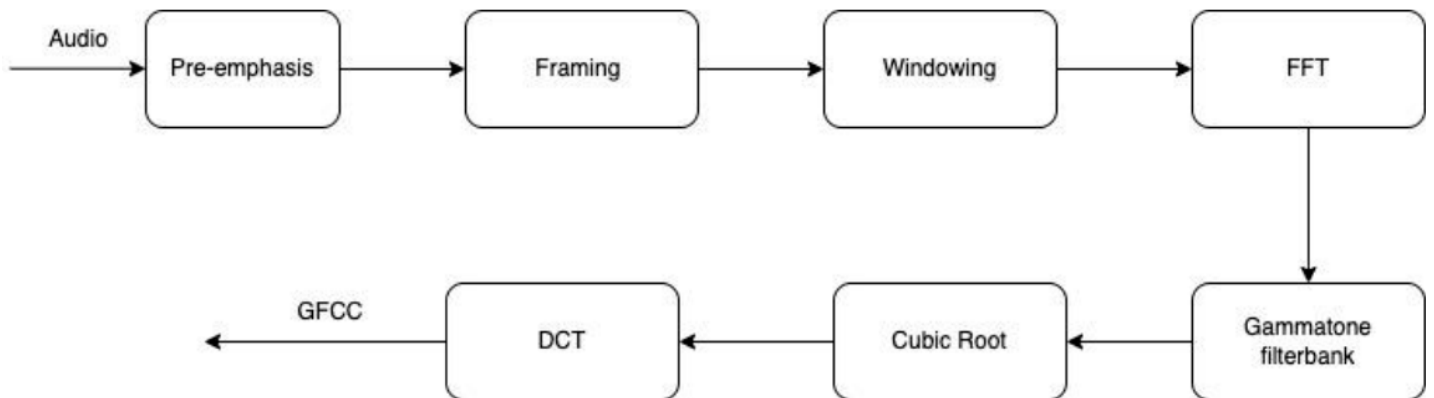


FIGURE 1. GFCC Feature Extraction Block Diagram

#### Pre-emphasis

This is the initial stage which functions to filter human speech signals. In the speech signal processing, a pre-emphasis filter is essential after the sampling process [21]. The purpose of this filtering is to achieve a smoother spectral frequency shape of the speech signal. The spectral shape is relatively high for the low-frequency region and tends to decrease sharply for frequencies above 2000 Hz. The pre-emphasis filter is grounded in the input/output relationship within the time domain, as expressed in the Equation (1).

$$y(n) = x(n) - a \cdot x(n - 1) \quad (1)$$

Where  $a$  represents the pre-emphasis filter constant, typically falling within the range of  $0.9 < a < 1.0$

### Framing

At this stage the signal is divided into several frames having a shorter duration because the signal in speech is always changing due to a shift in the articulation of the organs that reproduce sound. The frame size must be as long as possible, but it must also be short enough to obtain good time resolution. The frame length, often referred to as the window size, is usually chosen based on the characteristics of the speech signal and the requirements of the processing algorithm. Common frame lengths range from 20 to 30 milliseconds, but the optimal length can vary depending on the specific application. In this work, we decide frame length 30ms. This process is conducted in an overlapping manner to avoid loss of characteristics in each piece of the frame. For the overlap area, it covers 30 % and is carried out for each frame until all signals have been processed [22].

### Windowing

Due to the frame blocking (framing) process, the signal to be discontinued [22]. Thus, the windowing process aims to minimize signal discontinuities at the beginning and end of each frame. Windowing is performed frame by frame, and for each frame, a specific window function is applied. If we define the window as  $w(n)$ , where  $0 \leq n \leq N-1$ , with  $N$  being the number of samples in each frame, then the result of windowing is expressed as Equation (2).

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

### Fast Fourier Transform (FFT)

FFT is a method for converting sound signals into spectral components which provides frequency information about the sound signal. FFT is an algorithm for efficiently computing the Discrete Fourier Transform (DFT) with the aim of reducing digital calculations to simplify the calculation of the frequency spectrum in its implementation [23]. This method implements an algorithm that operates on discrete signals. FFT formula for  $N$  samples is described in the Equation (3).

$$x(n) = \sum_{k=0}^{N-1} x_k e^{\frac{-2\pi jkn}{N}} \quad (3)$$

Where  $0 \leq n \leq N-1$  and  $j = \sqrt{-1}$ .

### Gammatone Filter banks

This method is a form of imitating the cochlea of the human ear in processing sound. In this process, the sound signal that has been processed using the FFT becomes the frequency domain into the time-frequency domain [24]. The Gammatone filter bank is a set of filters commonly used in auditory processing models to simulate the frequency analysis performed by the human auditory system. Each filter in the bank is designed to mimic the response of the human auditory nerve fibers to different frequency components of a sound signal.

The Gammatone filter response is characterized by a shape that resembles a gamma distribution. The transfer function for a single Gammatone filter can be expressed as Equation (4).

$$g(t) = a \cdot t^{n-1} e^{-2\pi bt} \cos(2\pi f_c t + \phi) \quad (4)$$

Where  $f$  (in Hz) is the center frequency,  $\phi$  (in radians) is the phase of the carrier,  $a$  is the amplitude,  $n$  is the filter's order (set to 4),  $b$  (in Hz, set to 1.019) is the filter's bandwidth, and  $t=1/(2\pi.ERB(f))$  (in seconds) is the filter time constant, and  $ERB(f_c)$  is the equivalent rectangular bandwidth function, expressed as Equation (5).

$$ERB(f_c) = 24.7 \left( 4.37 \frac{f_c}{1000} + 1 \right) \quad (5)$$

### Cubic Root

Diverging from the logarithmic operations utilized in MFCC, the proposed GFCC employed cubic roots to compress the results of the filter bank [12]. The cubic roots better approximate the nonlinear response of the human auditory system to sound intensity variations, making them more relevant in representing auditory perception [25]. This compression technique enhances the discriminative power of the extracted features, particularly in tasks such as speech or sound classification, by capturing subtle differences in spectral characteristics [26]. The cubic root compression offers improved sensitivity to low-level details in the signal, crucial for distinguishing between similar sounds, as well as reducing the sensitivity to extreme values, leading to more robust feature extraction.

In addition, we mitigate the effects of temporal variations in the input signal by down sampling process, thereby improving the robustness of the feature representation to changes in signal duration. Finally, the equation of the operation cubic root expressed in Equation (6):

$$G_m[i] = |g_{downSampled}[i, m]|^{\frac{1}{3}}, i = 0 \dots N - 1, m = 0 \dots M - 1 \quad (6)$$

### Discrete Cosine Transform (DCT)

Finally, the results of the previous stage are processed using DCT. The aim of the final stage is to do decorrelation and reduce the dimensions of the features that have been produced. Unlike the Discrete Fourier Transform (DFT), which uses complex exponentials, the DCT uses only real numbers, making it computationally more efficient [27]. The Discrete Cosine Transform for the one-dimensional DCT of a sequence  $x[n]$  of length  $N$  is given in Equation (7).

$$X[k] = \sqrt{\frac{2}{N}} \cdot C(k) \cdot \sum_{n=0}^{N-1} x[n] \cdot \cos\left(\frac{\pi(2n+1)k}{2N}\right) \quad (7)$$

where  $X[k]$  is the DCT coefficient at frequency index  $k$ .  $C(k)$  is a scaling factor given by  $C(k)=1/\sqrt{2}$  for  $k=0$  and  $C(k)=1$  for  $k > 0$ .  $n$  is the time or spatial index.  $k$  is the frequency index.

### Random Forest (RF) Classifier

RF is a combination of decision tree algorithms that are considered weak in estimating, so RF combines decision trees to make stronger forecasts [28]. RF are more popular than other machine learning algorithms in over the last two decades because it can handle outliers and noisier datasets well; it also has higher accuracy and good performance with high dimensional datasets; another factor is that RF only requires two parameters  $n_{tree}$  and  $m_{try}$  to be optimized [29].



## MATERIALS AND METHODS

During the design stage of the system, the sound dataset was extracted using GFCC method. The extracted results then become input for Random Forest Classifier (RFC) and was divided into test data and training data. The RFC processed the sound features through 100 decision trees, which were then used as models and stored in “.pkl” format.

The mobile application design consists of use case design, flow diagram, and wireframe. The created mobile application was designed to have numbers of main functions such as connecting the application with the Raspberry Pi 4 device via Bluetooth, allowing the user to select the level of emotion to be detected, sending commands to the Raspberry Pi 4 to record sound, and receiving messages from the Raspberry Pi 4 to display the level results of the selected emotions.

The implementation of the previously designed system begins with deploying the program code and the model into Raspberry Pi 4, followed by making a Bluetooth connection by checking the Bluetooth address of the Raspberry Pi 4 and adding port 22. Furthermore, the mobile application implementation includes the steps of making block diagrams and interface design using MIT App Inventor, where block diagrams are structured using simple programming logic and application views are organized in sections Designer <sup>[30]</sup>. After all the systems have been set up, we did the preparation prototype tool by connecting microphone to Raspberry Pi 4 via port USB and connect the Raspberry Pi 4 to a power source portable form power bank.

For SNR testing, we tested the samples using both GFCC and MFCC extraction. GFCC accuracy test was done by comparing input sound received via microphone by modelling the emotional level of fear stored in the system's memory, then classifying the sound input using Random Forest Classifier to determine the appropriate emotion level of fear. Classification results were then sent to the mobile application via a Bluetooth connection and displayed on the app so that it can be seen by the user.

### Design and Implementation

#### *Sound Recording Hardware Design*

In the design of subsystem for recording sound, two pieces of hardware are required, microphone and the Raspberry Pi 4B. Sound signal acquired by the USB microphone will be recorded and stored in Raspberry Pi 4B memory in wav format. Figure 2 illustrates the voice acquisition subsystem.

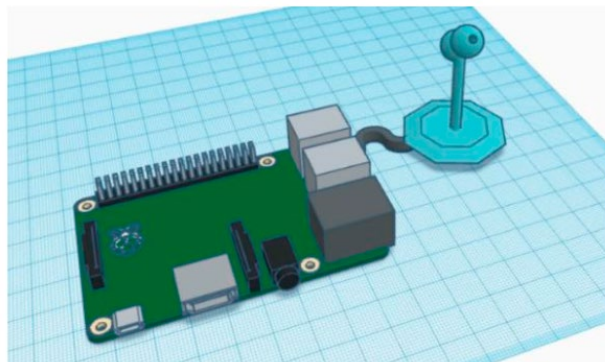


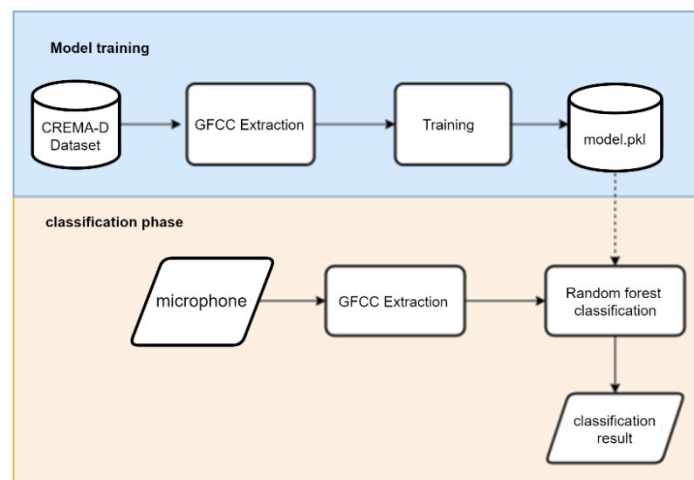
FIGURE 2. Sound Recorder Hardware Design

### Dataset

Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) is one of emotion dataset of 7,442 original clips featuring performances by 91 actors [31]. These actors, consisting of 48 males and 43 females spanning ages 20 to 74, represent diverse races and ethnicities, including African American, Asian, Caucasian, Hispanic, and Unspecified. The dataset includes recordings of actors delivering 12 sentences, each portraying one of six distinct emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) at four different emotion levels (Low, Medium, High, and Unspecified). In this work, we use only fear data of CREMA-D that consist of 273 utterances with 91 utterances for each class (Low, Medium, High).

### Sound Extraction & Classification Design

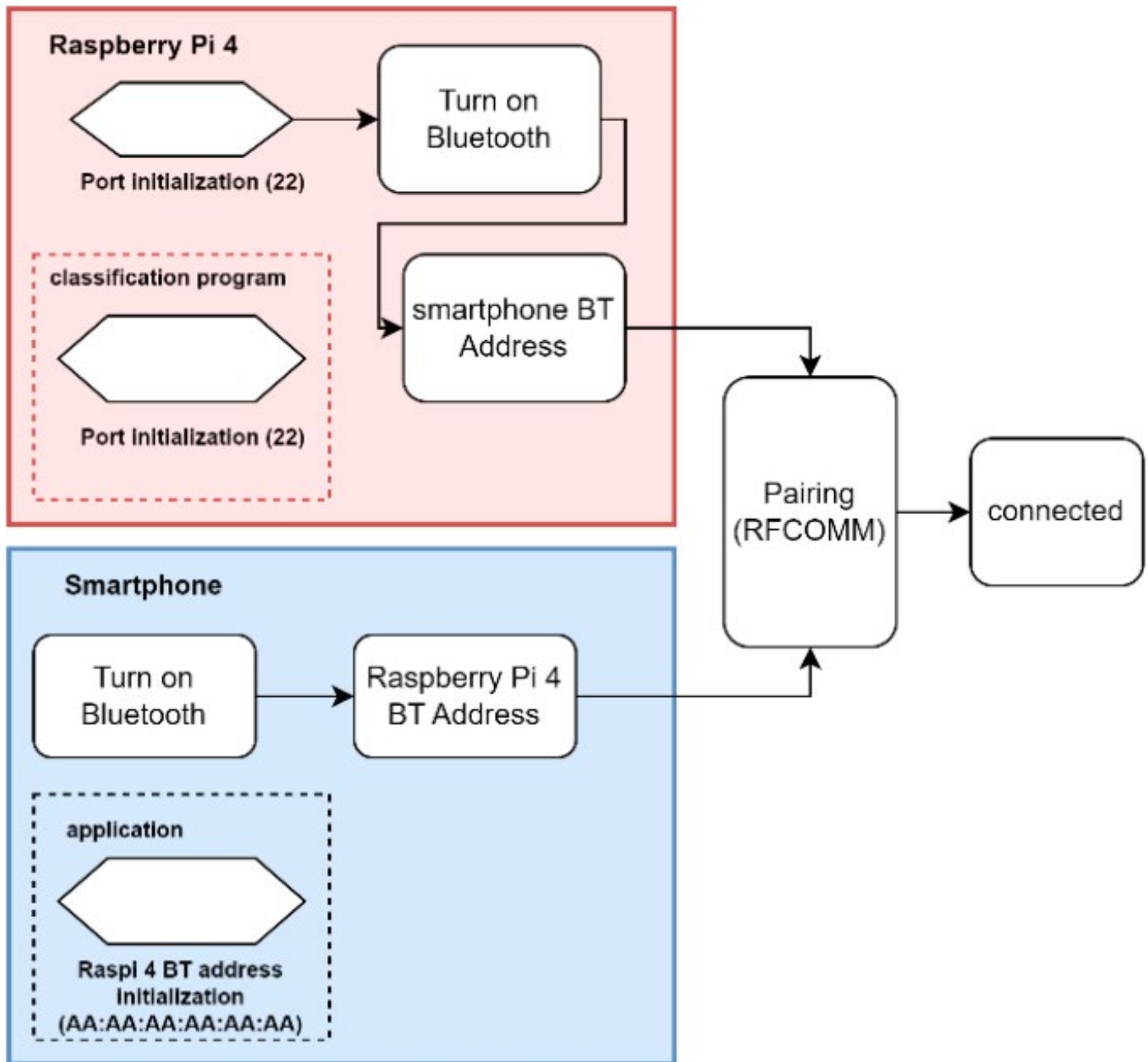
Initially, CREMA-D dataset consist of three levels of fear emotions, HI (high), MD (medium), and LO (low). Each utterance was then extracted using the GFCC method to identify the characteristics of each sound through its amplitude, frequency, and pattern. After that, the sound extraction results become the input for classification using the Random Forest (RF) Classifier. The extraction method with the highest accuracy was the selected to be implemented into the developed system. The results of the extraction and classification produce a model in “pkl” format. Figure 3 shows the flowchart of feature extraction and classification steps.



**FIGURE 3. Extraction and Classification Flowchart Design**

### Bluetooth Connection Design

Based on the system requirements it was decided to use the RFCOMM Bluetooth protocol to connect Raspberry Pi 4 to smartphone. The Bluetooth RFCOMM (Radio Frequency Communications) protocol is one of the protocols used in Bluetooth technology to govern communication between Bluetooth devices. The RFCOMM protocol provides a logical channel that can be used to send and receive data between Bluetooth devices and provides several additional services such as device recognition services and authentication services. Figure 4 shows the flow diagram of Bluetooth connection between devices.



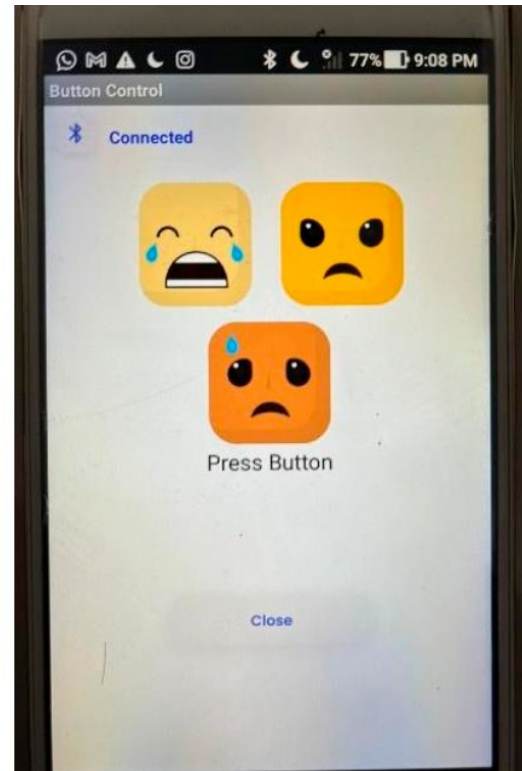
**FIGURE 4. Bluetooth Connection Flowchart**  
*System Implementation*

The System implementation was carried out based on the previous design phase. Initially, the first step was to prepare the box as a container. The microphone was connected through the USB port of the Raspberry Pi 4. The power source used to deliver the power for the Raspberry Pi 4 was obtained from a power bank whose cable was modified by adding a switch attached to the box so that the users do not need to open the box if they want to turn on the system. Apart from that, a hole was also made at the top of the box to become a place for the microphone. After all the devices had been set up, the housing box was closed and covered so that the system became portable device. The classification algorithm was implemented in Raspberry Pi, while the smartphone was used mainly to

display the result via bluetooth. Figure 5 depicts the overall implemented system as a portable device. Apart from hardware, the implementation of mobile application can be seen in Figure 6.



**FIGURE 5. Prototype of the System as portable device**



**FIGURE 6. Mobile Application User Interface**

In contrast to Mishra, *et al.* [32] and Alshamsi's, *et al.* [33] utilization of single hardware, such as Raspberry Pi or smartphones, for their emotion recognition systems, our approach involves the simultaneous integration of two hardware components - the Raspberry Pi and a smartphone. This strategic is rooted in the advantageous future utilization of both devices, facilitating the more effective incorporation of multimodal sensor inputs. The Raspberry Pi interfaces seamlessly with a diverse array of sensors, including microphones, cameras, and physiological sensors, enabling the capture of varied data modalities associated with emotions, such as speech, facial expressions, and physiological signals. This approach offers several notable benefits. Firstly, it allows for the collection of multimodal data, enriching the understanding of emotions. Secondly, the computational capabilities of the Raspberry Pi support initial data processing and feature extraction, complemented by the smartphones' enhanced processing power for advanced analysis and interaction. Additionally, the combined portability and affordability of the Raspberry Pi and smartphones enhance the system's versatility, enabling its use in diverse or multiple environments.

## RESULTS AND DISCUSSION

The effectiveness of the proposed system is evaluated in term of Signal to Noise Ratio (SNR) and accuracy. In all experiments, we divide the CREMA-D dataset into around 5 % (15 utterances) as test data, 15% as data validation (40), and 80 % as training data (218 utterances). The 15-testing data are selected randomly with different subject and level of fear emotion.

### Signal Testing-to-Noise Ratio (SNR)

Signal-to-noise ratio (SNR) is the ratio between the desired signal and noise in a system or signal. SNR can also be calculated by calculating the ratio between the amplitude of the desired signal to the amplitude of the noise.

$$SNR = 10 \left( \frac{S}{N} \right) dB \quad (8)$$

This test aims to determine which extraction has the highest SNR value between Raw, GFCC and MFCC. The test results are shown in Table 1.

**TABLE 1. SNR Test Results.**

| Sample*         | Signal-to-Noise Ratio (dB) |        |        |
|-----------------|----------------------------|--------|--------|
|                 | Raw                        | GFCC   | MFCC   |
| 1001-IEO-FEA-HI | 52.45                      | 174.50 | 78.46  |
| 1002-IEO-FEA-HI | 59.18                      | 160.69 | 93.90  |
| 1005-IEO-FEA-HI | 52.30                      | 74.60  | 106.42 |
| 1008-IEO-FEA-HI | 51.14                      | 149.82 | 89.07  |
| 1011-IEO-FEA-HI | 52.53                      | 96.85  | 89.95  |
| 1001-IEO-FEA-MD | 45.00                      | 184.36 | 86.32  |
| 1004-IEO-FEA-MD | 61.97                      | 238.64 | 92.62  |
| 1005-IEO-FEA-MD | 54.42                      | 101.56 | 97.04  |
| 1007-IEO-FEA-MD | 50.84                      | 138.47 | 89.17  |
| 1008-IEO-FEA-MD | 48.99                      | 125.39 | 91.36  |
| 1001-IEO-FEA-LO | 49.68                      | 124.94 | 85.42  |
| 1002-IEO-FEA-LO | 58.02                      | 69.44  | 101.27 |
| 1003-IEO-FEA-LO | 60.47                      | 75.20  | 97.71  |
| 1036-IEO-FEA-LO | 50.67                      | 165.56 | 102.25 |
| 1039-IEO-FEA-LO | 60.64                      | 128.06 | 106.39 |

\* Sample code: the first number indicate subject number

Based on the test results, 12 out of 15 or 80 % of the data samples had a higher SNR using GFCC extraction. From the table of SNR test results, the average SNR results for each label are presented in Table 2. It can be seen in the table that the highest SNR was obtained for sound data samples labelled "MID" using GFCC extraction with an average of 133.85 dB.

**TABLE 2. Average SNR Test Results for Each Label.**

| Label          | Signal-to-Noise Ratio (dB) |               |              |
|----------------|----------------------------|---------------|--------------|
|                | Raw                        | GFCC          | MFCC         |
| <i>HIGH</i>    | 53.52                      | 131.29        | 91.56        |
| <i>MEDIUM</i>  | 52.24                      | 157.64        | 91.30        |
| <i>LOW</i>     | 55.89                      | 112.64        | 98.60        |
| <b>Average</b> | <b>53.88</b>               | <b>133.85</b> | <b>93.82</b> |

### Accuracy

The aim of this test is to determine the accuracy of using GFCC extracted features as input and RFC classification in the emotion level detection system of fear. To obtain these results Equation (9) was used. The test results can be seen in Table 3. Based on these data it is known that the system succeeds in predicting correctly 11 out of 15 data samples so that the obtained accuracy was 73.33 %.

$$accuracy(\%) = \frac{\text{Total data is correct}}{\text{Total data}} \times 100 \quad (9)$$

**TABLE 3. Results of the Detection Accuracy of Fear Emotional Levels.**

| Sample*         | Label | Detection Results | Information |
|-----------------|-------|-------------------|-------------|
| 1008-IEO-FEA-HI | HIGH  | HIGH              | Correct     |
| 1034-IEO-FEA-HI | HIGH  | HIGH              | Correct     |
| 1048-IEO-FEA-HI | HIGH  | HIGH              | Correct     |
| 1025-IEO-FEA-HI | HIGH  | MID               | Wrong       |
| 1039-IEO-FEA-HI | HIGH  | HIGH              | Correct     |
| 1006-IEO-FEA-MD | MID   | MID               | Correct     |
| 1009-IEO-FEA-MD | MID   | MID               | Correct     |
| 1007-IEO-FEA-MD | MID   | MID               | Correct     |
| 1018-IEO-FEA-MD | MID   | MID               | Correct     |
| 1024-IEO-FEA-MD | MID   | MID               | Correct     |
| 1005-IEO-FEA-LO | LOW   | LOW               | Correct     |
| 1016-IEO-FEA-LO | LOW   | LOW               | Correct     |
| 1018-IEO-FEA-LO | LOW   | MID               | Wrong       |
| 1026-IEO-FEA-LO | LOW   | MID               | Wrong       |
| 1037-IEO-FEA-LO | LOW   | MID               | Wrong       |

\* Sample code: the first number indicate subject number

In addition, to ensure the effectiveness of the proposed system, we compare the proposed system with baseline system in term of SNR and accuracy. The comparison results show in Table 4.

**TABLE 4. Results of the Detection Accuracy of Fear Emotional Levels.**

| Reference                             | Number of Classes                                | Number of Features                                   | Accuracy |
|---------------------------------------|--|--|----------|
| Chebbi, <i>et al.</i> <sup>[34]</sup> | 3 classes (fear, neutral, other emotions)        | 11 (based on FDR and ANOVA)                          | 78 %     |
|                                       |  | 10 (based on Scatter measure)                        | 77 %     |
|                                       |  | 23 (based on divergence)                             | 83 %     |
|                                       |  | 8 (based on FDR, ANOVA, Scatter measure, Divergence) | 86 %     |
| Clavel, <i>et al.</i> <sup>[35]</sup> | 2 classes (fear and neutral)                     | -  | 70 %     |
| This work                             | 3 classes (high fear, medium fear, and low fear) | GFCC   | 73 %     |

Table 4 demonstrates the result performance of the proposed system compared baseline system. Chebbi, *et al.* [34] and Clavel, *et al.* [35] explored a range of features and achieved accuracies up to 86 % in a three-class fear classification task, while our work focuses on a more nuanced fear analysis, classifying fear into high, medium, and low levels. Despite achieving an accuracy of 73 %, which may be lower than the maximum reported by Chebbi, *et al.*, the multiclass approach adds granularity to the classification, providing valuable insights into different fear intensities. The use of GFCC as features in this work offers a distinct representation of the data, potentially capturing nuanced aspects of fear expression. Furthermore, the decomposition of fear levels allows for a more detailed understanding of emotional states, catering to specific applications where finer distinctions in fear intensity are essential. Overall, the advantages of this work lie in its multiclass approach, detailed fear level analysis, and the potential relevance of GFCC features in capturing the nuances of fear expressions.

## CONCLUSION

GFCC (Gamma Frequency Cepstral Coefficients) are better than MFCC (Mel-Frequency Cepstral Coefficients) in reducing noise in feature extraction of sound. This is due to the use of the gamma filter on the GFCC which can reduce noise. In the developed system, the fear emotion labelled with the "MID" label has the highest SNR (Signal-to-Noise Ratio) value by using GFCC. The GFCC extraction method and the Random Forest Classifier are quite effective for detecting the moderate emotional level of fear (MID) with an accuracy of around 73 %.

For future research directions, several areas have been identified. First, further optimization of the fear emotion recognition model can be pursued by considering various neural network architectures, activation functions, and training parameters to enhance accuracy and system responsiveness. Second, expanding the dataset by collecting additional data from diverse sources could improve the diversity and representation of fear expression variations. Exploring other frequency features to deepen the understanding and detection of fear emotion levels is another avenue for research. The selection and comparison with alternative machine learning techniques, such as Support Vector Machines (SVM) or Decision Trees, can also be a focus for future research. The integration of multimodal data, such as audio and visual information, and the addition of sentiment and context analysis elements can enrich the understanding of emotions. Furthermore, evaluating the system's effectiveness in real-world scenarios can provide richer insights and test the system's performance in complex and variable situations. Developments in these various aspects are expected to enhance the performance and usability of the fear emotion level recognition system based on GFCC.

## ACKNOWLEDGMENTS

We thank our colleagues from Embedded System and Robotics Laboratory Universitas Brawijaya that has supported us in this work.

## AUTHOR CONTRIBUTIONS

B. H. P. Conceptualization, methodology, validation, supervision, project administration, funding acquisition. L. O. A. H. Software, formal analysis, investigation, resources, data curation, writing original draft. D. S. Validation, formal analysis, data curation, writing review and editing, visualization. E. R. W. Methodology, validation, formal analysis, investigation, data curation, writing review and editing.

## REFERENCES

- [1] M. Gupta, S. S. Bharti, and S. Agarwal, "Gender-based speaker recognition from speech signals using GMM model," *Mod. Phys. Lett. B*, vol. 33, no. 35, 2019, doi: <https://doi.org/10.1142/S0217984919504384>
- [2] J. Otašević and B. Otašević, "Voice-based identification and contribution to the efficiency of criminal proceedings," *J. Crim. Crim. Law*, vol. 59, no. 2, pp. 61-72, Nov. 2021, doi: <https://doi.org/10.47152/rkkp.59.2.4>
- [3] S. A. Kotz, R. Dengler, and M. Wittfoth, "Valence-specific conflict moderation in the dorso-medial PFC and the caudate head in emotional speech," *Soc. Cogn. Affect Neurosci.*, vol. 10, no. 2, 2015, doi: <https://doi.org/10.1093/scan/nsu021>
- [4] S. J. Gomez, "Self-Management Skills of Management Graduates," *Int. J. Res. Manag. Bus. Stud.*, vol. 4, no. 3, pp. 40-44, 2017.
- [5] S. A. Saddiqui, M. Jawad, M. Naz, and G. S. Khan Niazi, "Emotional intelligence and managerial effectiveness," *RIC*, vol. 4, no. 1, pp. 99-130, 2018, doi: <https://doi.org/10.32728/RIC.2018.41%2F5>
- [6] H. E. Erskine, S. J. Blondell, M. E. Enright, J. Shadid, et al., "Measuring the Prevalence of Mental Disorders in Adolescents in Kenya, Indonesia, and Vietnam: Study Protocol for the National Adolescent Mental Health Surveys," *J. Adolesc. Health*, vol. 72, no. 1, pp. S71-S78, 2023, doi: <https://doi.org/10.1016/j.jadohealth.2021.05.012>
- [7] K. Cherry, "What Are Emotions and the Types of Emotional Responses?" *Verywell Health*. <https://www.verywellhealth.com/what-are-emotions-279517807> (accessed 2023).
- [8] S. Sharma, A. Mamata, Deepak, "Psychological Impacts, Hand Hygiene Practices & and Its Correlates in View of Covid-19 among Health Care Professionals in Northern States of India," *Indian J. Forensic Med. Toxicol.*, vol. 15, no. 2, pp. 3691-3698, 2021, doi: <https://doi.org/10.37506/ijfmt.v15i2.14947>
- [9] S. A. Mahar, M. H. Mahar, J. A. Mahar, M. Masud, M. Ahmad, N. Z. Jhanhi, and M. A. Razzaq, "Superposition of functional contours based prosodic feature extraction for speech processing," *Intell. Autom. Soft Comput.*, vol. 29, no. 1, pp. 183-197, 2021, doi: <https://doi.org/10.32604/iasc.2021.015755>
- [10] S. Sondhi, M. Khan, R. Vijay, A. K. Salhan, and S. Chouhan, "Acoustic analysis of speech under stress," *Int. J. Bioinform. Res. Appl.*, vol. 11, no. 5, pp. 417-432, 2015, doi: <https://doi.org/10.1504/ijbra.2015.071942>
- [11] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC," in 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 2017, pp. 2257-2260, doi: <https://doi.org/10.1109/WISPNET.2017.8300161>
- [12] M. Jeevan, A. Dhingra, M. Hanmandlu, and B. K. Panigrahi, "Robust speaker verification using GFCC based i-vectors," in Proceedings of the International Conference on Signal, Networks, Computing, and Systems. Lecture Notes in Electrical Engineering, vol 395. New Delhi, India, pp. 85-91, 2017, doi: [https://doi.org/10.1007/978-81-322-3592-7\\_9](https://doi.org/10.1007/978-81-322-3592-7_9)
- [13] H. Wang and C. Zhang, "The application of Gammatone frequency cepstral coefficients for forensic voice comparison under noisy conditions," *Aust. J. Forensic Sci.*, vol. 52, no. 5, pp. 553-568, 2020, doi: <https://doi.org/10.1080/00450618.2019.1584830>
- [14] D. Bharti and P. Kukana, "A Hybrid Machine Learning Model for Emotion Recognition from Speech Signals," in 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 491-496, doi: <https://doi.org/10.1109/ICOSEC49089.2020.9215376>
- [15] H. Patni, A. Jagtap, V. Bhojar, and A. Gupta, "Speech Emotion Recognition using MFCC, GFCC, Chromagram and RMSE features," in 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2021, pp. 892-897, doi: <https://doi.org/10.1109/SPIN52536.2021.9566046>
- [16] H. Choudhary, D. Sadhya, and V. Patel, "Automatic Speaker Verification using Gammatone Frequency Cepstral Coefficients," in 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2021, pp. 424-428, doi: <https://doi.org/10.1109/SPIN52536.2021.9566150>
- [17] L. Zheng, Q. Li, H. Ban, and S. Liu, "Speech emotion recognition based on convolution neural network combined with random forest," in 2018 Chinese Control And Decision Conference (CCDC), Shenyang, China, 2018, pp. 4143-4147, doi: <https://doi.org/10.1109/CCDC.2018.8407844>
- [18] S. Hamsa, I. Shahin, Y. Iraqi, and N. Werghe, "Emotion Recognition from Speech Using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier," *IEEE Access*, vol. 8, pp. 96994-97006, 2020, doi: <https://doi.org/10.1109/ACCESS.2020.2991811>
- [19] A. Cuncic, "Amygdala Hijack and the Fight or Flight Response," *Very Well Mind*. <https://www.verywellmind.com/what-happens-during-an-amygdala-hijack-4165944> (accessed 2023).
- [20] E. B. Foa and M. J. Kozak, "Emotional Processing of Fear. Exposure to Corrective Information," *Psychol. Bull.*, vol. 99, no. 1, pp. 20-35, 1986, doi: <https://psycnet.apa.org/doi/10.1037/0033-2909.99.1.20>
- [21] S. M. Qaisar, "Isolated speech recognition and its transformation in visual signs," *J. Electr. Eng. Technol.*, vol. 14, no. 2, pp. 955-964, 2019, doi: <https://doi.org/10.1007/s42835-018-00071-z>
- [22] S. Lokesh and M. R. Devi, "Speech recognition system using enhanced mel frequency cepstral coefficient with windowing and framing method," *Cluster Comput.*, vol. 22, pp. 11669-11679, 2019, doi: <https://doi.org/10.1007/s10586-017-1447-6>
- [23] J. D. Schmidt, "Simple Computations Using Fourier Transforms," in *Numerical Simulation of Optical Wave Propagation with Examples in MATLAB*, Bellingham, WA, USA: SPIE Press, 2010, doi: <https://doi.org/10.1117/3.866274.ch3>



- [24] A. Krobba, M. Debyeche, and S. A. Selouani, "Mixture linear prediction Gammatone Cepstral features for robust speaker verification under transmission channel noise," *Multimed. Tools Appl.*, vol. 79, no. 25-26, pp. 18679-18693, 2020, doi: <https://doi.org/10.1007/s11042-020-08748-2>
- [25] A. Revathi, N. Sasikaladevi, R. Nagakrishnan, and C. Jeyalakshmi, "Robust emotion recognition from speech: Gamma tone features and models," *Int. J. Speech Technol.*, vol. 21, no. 3, pp. 723-739, 2018, doi: <https://doi.org/10.1007/s10772-018-9546-1>
- [26] U. Kumaran, S. Radha Rammohan, S. M. Nagarajan, and A. Prathik, "Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN," *Int. J. Speech Technol.*, vol. 24, no. 2, pp. 303-314, 2021, doi: <https://doi.org/10.1007/s10772-020-09792-x>
- [27] S. Rhee, M. G. Kang, "Discrete cosine transform based regularized high-resolution image reconstruction algorithm," *Opt. Eng.*, vol. 38, no. 8, pp. 1348-1356, 1999, doi: <https://doi.org/10.1117/1.602177>
- [28] A. Subudhi, M. Dash, and S. Sabut, "Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier," *Biocybern. Biomed. Eng.*, vol. 40, no. 1, pp. 277-289, 2020, doi: <https://doi.org/10.1016/j.bbe.2019.04.004>
- [29] T. N. Phan, V. Kuch, and L. W. Lehnert, "Land cover classification using google earth engine and random forest classifier-the role of image composition," *Remote Sens.*, vol. 12, no. 15, art. no. 2411, 2020, doi: <https://doi.org/10.3390/rs12152411>
- [30] T. Adiono, S. F. Anindya, S. Fuada, K. Afifah, and I. G. Purwanda, "Efficient Android Software Development Using MIT App Inventor 2 for Bluetooth-Based Smart Home," *Wireless Pers. Commun.*, vol. 105, pp. 233-256, 2019, doi: <https://doi.org/10.1007/s11277-018-6110-x>
- [31] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377-390, 2014, doi: <https://doi.org/10.1109/taffc.2014.2336244>
- [32] A. Mishra, D. Patil, N. Karkhanis, V. Gaikar, and K. Wani, "Real time emotion detection from speech using Raspberry Pi 3," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, India, 2017, pp. 2300-2303, doi: <https://doi.org/10.1109/WiSPNET.2017.8300170>
- [33] H. Alshamsi, V. Kepuska, H. Alshamsi, and H. Meng, "Automated Speech Emotion Recognition on Smart Phones," in *2018 9th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, New York, NY, USA, 2018, pp. 44-50, doi: <https://doi.org/10.1109/UEMCON.2018.8796594>
- [34] S. Chebbi and S. Ben Jebara, "On the Selection of Relevant Features for Fear Emotion Detection from Speech," in *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*, Rabat, Morocco, 2018, pp. 82-86, doi: <https://doi.org/10.1109/ISIVC.2018.8709233>
- [35] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Commun.*, vol. 50, no. 6, pp. 487-503, 2008, doi: <https://doi.org/10.1016/j.specom.2008.03.012>