Quest Journals
Journal of Software Engineering and Simulation
Volume 12 ~ Issue 4 (April 2026) pp: 27-33
ISSN(Online) :2321-3795 ISSN (Print):2321-3809
www.questjournals.org

Research Paper

Advancing Emotion Recognition through Voice: A System Development Approach

Mohd Amzar Azizan[1,2], Fawad Ahmad[1,2], Anish Joy[1,2]
[1] *Higher College of Technology, Abu Dhabi, UAE*
[2] *Khalifa bin Zayed Air College, UAE*

ABSTRACT
In light of technological advancements, expanding speech, the most natural means of human expression, into computerized applications has become crucial. This study focuses on developing a medium for machines to internalize human emotions using the RAVDESS dataset as a reliable reference, and leveraging Artificial Neural Network (ANN) and Mel Frequency Cepstral Coefficients (MFCC) for analysis. The primary objective is to quantify the contributions of Artificial Intelligence (AI) and Machine Learning (ML) in recognizing human emotions through voice. The collected data will be transmitted to the cloud, aiming to provide assistance to emotionally-disturbed individuals. This research provides novel insights into the effective management of human emotions through the integration of AI and ML. By utilizing high-quality microphones with active noise-cancelling features, a user-friendly interface can be created to monitor emotional outcomes and deliver accurate results to users. The implementation of coding allows the system to automatically assess users' emotional states, eliminating the need for manual button-pushing. This is made possible by leveraging large datasets and supporting various recognized languages. The significance of this study lies in its direct relevance to human awareness, as uncontrolled emotions are strongly linked to depression, mental health issues, and suicide attempts. Future studies should consider incorporating a wide range of languages to ensure the recognition and appropriate consideration of expressed emotions, taking into account the implications of linguistic variations on emotional expressions. By bridging the gap between human emotions and machine applications, this research contributes to the development of technologies that can effectively manage and address emotional well-being

Keywords: *Emotions, RAVDESS; Internet of Things; MFCC; Human-Computer Interaction (HCI)*

*Received 28 Mar., 2026; **R**evised 06 Apr., 2026; **A**ccepted 08 Apr., 2026 © The author(s) 2026.*
P**ublished with open access at **www.questjournals.org

I. Introduction

Individuals have increasingly relied on machines and technologies to simplify their lives and perform daily chores [1]. However, the current-utilized equipment has certain drawbacks, such as the inability to internalize or interpret human moods and stress levels. The ability to sense and understand human emotions expressed through speech, bodily movements, and facial expressions is pivotal in interpersonal connections [2].

Emotion recognition in speech has emerged as a crucial aspect of human-machine communication and work quality. A voice emotion recognition system has been developed to acknowledge and interpret emotions based on users' speech input [1]. This system utilizes fundamental components that resemble those of conventional pattern recognition systems. Notably, the speed and pitch of speech differ between genders, and the frequency range of speech shifts with its amplitude [2]. In a specific study that utilized emotional speech as input, the measured pitch ranged between 90Hz and 200Hz [3][4].

Speech recognition systems, as a domain in computer science, recognize and interpret spoken utterances by digitizing their sounds and comparing them with patterns from past and current data. Speech emotion detection primarily aims to determine the appropriate emotions for voice emotion processing, given their significance in conveying individual viewpoints and emotional states [5][6].

In the current study, algorithms are presented to provide readers with a comprehensive understanding of the examined topic. These algorithms are based on a neural network model inspired by the human brain. Building upon the research conducted by D. Doye et al. in 2002 [7], the equipment used in this study was selected based on the eyes and tailored for a specified application through a learning process. The study highlights three primary

components for speech emotion recognition: (i) emotional speech database selection, (ii) audio data feature selection, and (iii) classifier selection [8][9]. One notable and accurate database used in this field is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), which encompasses emotional speech and songs and is compatible with multi-modal analysis [10].

The objectives of this study are to develop a device that can comprehend human emotions through speech and design an algorithm utilizing the Mel Frequency Cepstral Coefficient (MFCC) for speech-based emotion recognition [11].

II. Methodology

This section discusses the steps undertaken in the current study, including software, hardware, and data collection and analysis. Voice features such as MFCC and Mel spectrograms play a crucial role in identifying speakers, recognizing voices, and assessing emotional conditions in voice emotion detection [12]. The MFCC is commonly used in speech and voice emotion recognition models due to its consistent and accurate performance even with variations in voice pitch [13][14][15].

The data-gathering procedure involved passionate speech into a microphone connected to a Raspberry Pi 4. The program was trained using datasets prior to its application on the Raspberry Pi 4, which demonstrated system accuracy. The collected data were then tabulated to establish the validity of emotion provided by the equipment. The outcomes of the experiment will be presented in the following section.

2.1 Equipment

2.1.1 Raspberry Pi 4 The Raspberry Pi 4 is a single-board computer with a compact footprint that can be used as a small computer by adding a keyboard, mouse, and display. This platform is widely utilized for real-time image or video processing, IoT applications, and robots. The Raspberry Pi Foundation has certified Raspbian OS as a free operating system specifically designed for Raspberry Pi. The Raspbian graphical user interface (GUI) serves as the central point for processing all data and producing output, providing tools for web browsing, Python programming, and office productivity

Fig. 1. Board of Raspberry Pi

2.1.2 Micro SD card

As the smallest consumer flash memory card in the world (approximately the diameter of the fingernail), a micro secure digital (SD) card contains the same electrical connectors as a normal counterpart. Micro SD cards may now be used in standard SD card slots through an adaptor: a removable flash memory card that is primarily used in mobile phones. Parallel to any other flash memory cards, the micro SD card could store various entities ranging from images, vidoes and music to software. In this study, the micro SD card was utilised to store both the Phyton code and datasets necessary for voice input comparison.

Fig. 2. Micro SD card

2.1.3 The USB Microphone

The high quality microphone with an 'integrated' interface that connects to the USB port on one's computer could record without relying on the built-in computer sound card for significantly enhanced outcomes. This microphone, which includes the necessary amplification to guarantee appropriate signal amplification, was selected following its compact size and lack of additional wires for operational purposes.

Fig. 3. Mini-USB microphone

2.1.4 The LCD Screen

A liquid-crystal display (LCD) denotes a flat-panel display or other optical devices that utilize liquid crystals and polarizers for light manipulation. Essentially, liquid crystals employ a backlight or reflector to create colour or monochrome images without direct light generation. The screen was connected to Raspberry Pi 4 to display the Raspberry Pi 4 processing results

Fig. 4. The LCD screen

2.1.5 Phyton

Following its compatibility with Raspberry Pi and user-friendliness, Python was incorporated as a programming language into this study given the abundance of easily-importable libraries for prototyping purpose: a no-brainer in machine learning.

Fig. 5. Python programming

III. Results

The accuracy and emotion recognition tests were conducted to evaluate the precision of user emotion detection and output. Additionally, the stress level was assessed after the display of emotions. This comprehensive test encompassed individuals of various genders and ages to determine the accuracy of different voices across diverse demographics. The Python programming language was employed to determine the dataset accuracy by comparing it with the training data. Interestingly, no significant differences in accuracy were observed for any of the programmed emotions. Figure 6 illustrates the libraries and models utilized in the experiment.

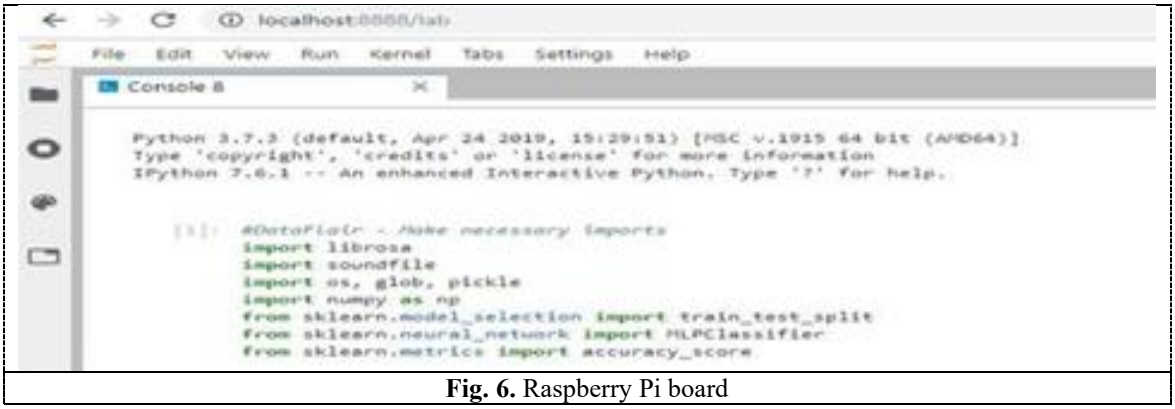

Fig. 6. Raspberry Pi board

3.1.1 The effect of angle of attack

Imported sound files involving MFCC, Mel and Chroma features were extracted and employed. The sound file was then extracted. The RAVDESS dataset must be categorised pre-extraction following Figure 7 and 8.

```
[2]: #DataFlair - Extract features (mfcc, chroma, mel) from a sound file
     def extract_feature(file_name, mfcc, chroma, mel):
         with soundfile.SoundFile(file_name) as sound_file:
             X = sound_file.read(dtype="float32")
             sample_rate=sound_file.samplerate
             if chroma:
                 stft=np.abs(librosa.stft(X))
             result=np.array([])
             if mfcc:
                 mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T, axis=0)
                 result=np.hstack((result, mfccs))
             if chroma:
                 chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
                 result=np.hstack((result, chroma))
             if mel:
                 mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
                 result=np.hstack((result, mel))
         return result
```

Fig. 7. Feature Extraction used

```
[3]: #DataFlair - Emotions in the RAVDESS dataset
     emotions={
         '01':'neutral',
         '02':'calm',
         '03':'happy',
         '04':'sad',
         '05':'angry',
         '06':'fearful',
         '07':'disgust',
         '08':'surprised'
     }

     #DataFlair - Emotions to observe
     observed_emotions=['calm', 'happy', 'fearful', 'disgust']
```

Fig. 8. Emotion Classification

The number denotes the feeling. The data were then extracted for each sound file with 'x' denoting features and 'y' implying emotions. The function train test split, test size, and random state value were all referred to as this. Figure 9 depicts the extract features for each sound file

```
[4]: #DataFlair - Load the data and extract features for each sound file
     def load_data(test_size=0.2):
         x,y=[],[]
         for file in glob.glob("D:\\DataFlair\\ravdess data\\Actor_*\\*.wav"):
             file_name=os.path.basename(file)
             emotion=emotions[file_name.split("-")[2]]
             if emotion not in observed_emotions:
                 continue
             feature=extract_feature(file, mfcc=True, chroma=True, mel=True)
             x.append(feature)
             y.append(emotion)
         return train_test_split(np.array(x), y, test_size=test_size, random_state=9)
```

Fig. 9. Extracted features for each sound file

At this stage, the dataset was divided into training and testing sets, with 75% allocated for training and 25% for evaluation, ensuring optimal data training. The test results provided a measure of emotion accuracy after the training process. Figure 10 illustrates the accuracy of the trained model.
To achieve accurate results, users were instructed to speak clearly and express their desired emotions. This instruction was applicable to male and female adults, teenagers, and children. Prior to voice recording, users were prompted with a question, such as "How are you?" to elicit their emotional response. Users then proceeded to speak, allowing their voices to be recorded, and the resulting emotional output to be transmitted to the cloud.

The process involved users pressing a button on the voice emotion recognition device to initiate voice recording using the Raspberry Pi 4-connected microphone. The recorded voice was then interpreted by a Python-trained dataset. Additionally, artificial intelligence (AI) was employed to recognize users' sentiments post-interpretation, and the output was subsequently transmitted to the cloud server.

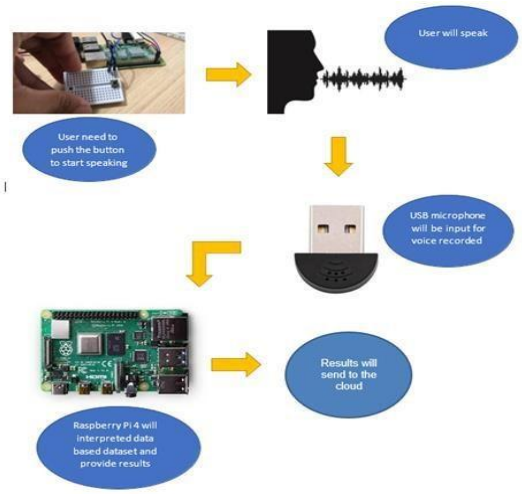

Fig. 10. System layout

Users were required to install the VNC viewer to connect the Raspberry Pi to the laptop and an ethernet or Lan cable to view the screen display. Figure 11 and 12 illustrate the device setup and VNC viewer apps.

Fig. 11. Device Setup

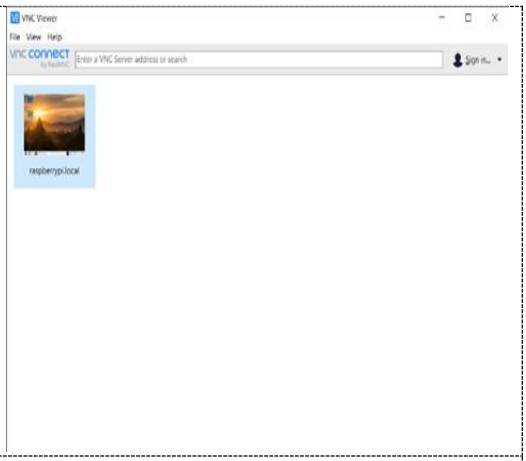

Fig. 12. The VNC Viewer Apps

To initiate testing, users were instructed to open the source code on the Raspberry Pi and run the script. The coding process indicated the machine's readiness to start, as shown in Figure 13. Next, users were required to push the switch button on the hardware to initiate voice emotion recognition, as depicted in Figure 14.
During the testing phase, users' voices were recorded for a duration of 15 seconds. After the recording, the emotion data was processed and presented, as illustrated in Figure 15. The recorded data, including the date and time, employee name and identification, and the current emotion during voice emotion recording, were then transmitted to the cloud server. Figure 16 showcases the documentation of this information

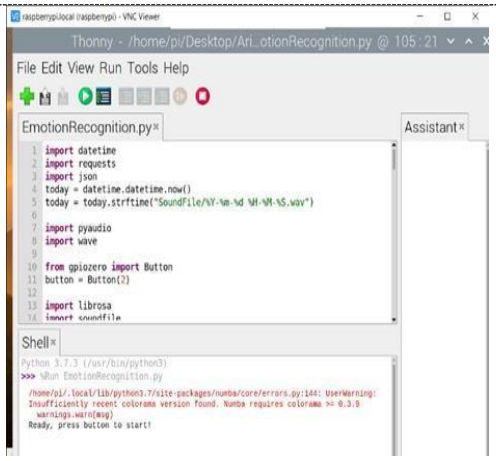

Fig. 13. Voice Emotion Ready To Use

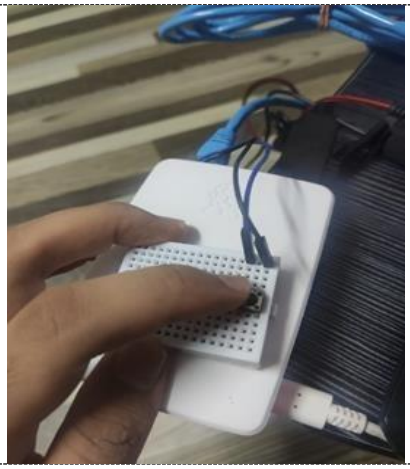

Fig. 14. Switch Button

Fig. 15. Results of Voice Recorded

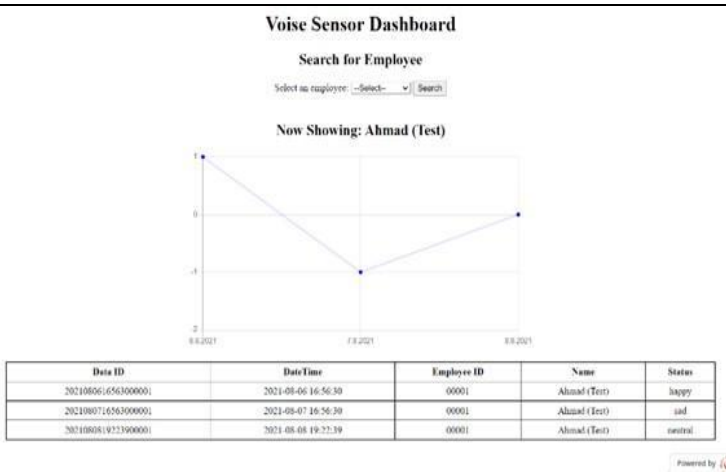

Fig. 16. User Emotion on Cloud Server

Figure 16 presents the voice emotion recognition cloud server dashboard whose design is based on employees where the manager could monitor their emotions with the toggle search in the dashboard and select other employee emotions stored in the cloud. This practice proves beneficial given the high number of stressed and depressed employees who are only identified when they are past help.

IV. Conclusions

In conclusion, voice emotion recognition technology has the potential to significantly benefit various work and educational settings, especially in the context of online learning methods. By accurately detecting users'

emotions through voice analysis, this technology can help establish a conducive environment that addresses mental health issues.

The impact of the COVID-19 pandemic, including prolonged periods spent at home and increased stress levels, has resulted in a higher number of suicide cases. In such circumstances, the implementation of highly accurate voice emotion recognition systems becomes crucial. These systems ensure that the derived outcomes precisely reflect users' emotional states, allowing for better understanding and support.

One notable advantage of the study mechanism employed in this research is its relative ease and cost-effectiveness compared to previous emotion recognition systems. Unlike systems that rely on high-quality cameras for emotion detection, the voice-based approach presents a more affordable alternative. However, it is important to acknowledge that the absence of visual cues may limit the system's ability to capture certain aspects of emotional expression.

To further advance the field of voice emotion recognition and improve the research, several recommendations can be considered. First, the integration of multi-modal data, such as facial expressions and body movements, alongside voice analysis, could provide a more comprehensive understanding of users' emotional states. Exploring advanced machine learning techniques, such as deep learning models, can also enhance the accuracy and robustness of the system.

Additionally, improving the user interface of the voice emotion recognition system to enhance user-friendliness and providing real-time feedback can enhance user engagement. Collaborating with mental health professionals can validate the efficacy and potential therapeutic applications of the system, tailoring it to specific counseling or therapeutic purposes.

In summary, voice emotion recognition technology holds promise for creating conducive environments, particularly in work and educational settings, to address mental health issues. While the current study mechanism offers advantages in terms of ease and affordability, future research should focus on integrating multi-modal data, exploring advanced machine learning techniques, improving the user interface, and collaborating with mental health professionals to maximize the potential of voice emotion recognition systems.

References

[1]. Lech M, Stolar M, Best C and Bolia R (2020) Real-Time Speech Emotion Recognition Using a Pretrained Image Classification Network: Effects of Bandwidth Reduction and Companding. *Front. Comput. Sci*. 2:14. doi: 10.3389/fcomp.2020.00014

[2]. I. C. Jang, M. K. Park, T. S. Kim and M. W. Park, "Study of emotion in speech", *Korean Society for Precision Engineering Conference*, pp. 25-28, October 2004

[3]. B. Schuller et. al HIDDEN MARKOV MODEL-BASED SPEECH EMOTION RECOGNITION 2003

[4]. C.H. Kwon and S.K. Song, "Extraction of speech features for emotion recognition", *Phonetics and Speech Sciences*, vol. 4, no. 2, pp. 73-78, Jun 2012.

[5]. Zhu, Wenjing and Li, Xiang Speech Emotion Recognition with Global-Aware Fusion on Multi-scale Feature Representation https://doi.org/10.48550/arXiv.2204.05571

[6]. Hadhami Aouani, Yassine Ben Ayed, Speech Emotion Recognition with deep learning, *Procedia Computer Science*, Volume 176, 2020, 251-260, https://doi.org/10.1016/j.procs.2020.08.027

[7]. D. Doye, U. Kulkarni, and T. Sontakke, "Modified Fuzzy Hypersphere Neural Network Recognition," *Proceedings of the International Joint Conference on Neural Networks*, vol. 1, pp. 65–68, 2002.

[8]. M.E.Ayadi, M.S.Kamel, F.Karray (2011). Features, categorization, and speech emotion recognition databases. Recognition pattern, vol. 44, 3, pp. DOI: 572–587. 10.1016/j.patcog.2010.09.020.

[9]. Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access*, 7, 117327–117345. doi:10.1109/access.2019.2936124

[10]. R. Anusha, P. Subhashini, D. Jyothi, P. Harshitha, J. Sushma and N. Mukesh, "Speech Emotion Recognition using Machine Learning," *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2021, pp. 1608-1612, doi: 10.1109/ICOEI51242.2021.9453028.

[11]. Selvaraj, Mahalakshmi & Bhuvana, R. & Karthik, S Padmaja. (2016). Human speech emotion recognition. 8. 311-323.

[12]. Han Wei et al., "An efficient MFCC extraction method in speech recognition", 2006

[13]. A.B. Ingale and D.S. Chaudhari, "Speech emotion recognition", *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 1, pp. 235-238, 2012

[14]. Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors* 2021, 21, 1249. https://doi.org/10.3390/s2104124

[15]. Albadr, M.A.A., Tiun, S., Ayob, M. *et al.* Speech emotion recognition using optimized genetic algorithm-extreme learning machine. *Multimed Tools Appl* 81, 23963–23989 (2022). https://doi.org/10.1007/s11042-022-12747-w