MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1189-1194

AN INTELLIGENT ROBOT FOR REAL-TIME VOICE EMOTION DETECTION AND EXPRESSIVE RESPONSE

M.Gomathi[1] , S.Ananth[2],K.Rahmaan[3] Alif Lam[4],
P.T. Shefi Munees Ramadan[5], Sreeram Dilly[6]
Assistant Professor[1], Associate Professor[2], Assistant Professor[3], UG student [4, 5, 6]
Department of Artificial Intelligence & Data Science, Mahendra Engineering College,
Mallasamudram, Namakkal, Tamil Nadu, - 637 503
m.gomath@gmail.com[1],ananths@mahendra.info[2],rahmaank@mahendra.info[3], mralif360@gmail.com[4], shefiptshefi@gmail.com[5], sreeramdilly@gmail.com[6]

ABSTRACT
This studies on voice-based human–robot interaction highlight the importance of integrating emotion recognition into intelligent systems to improve user experience. Existing voice assistants primarily focus on speech recognition and command execution, but they lack the ability to understand the emotional state of the user, resulting in limited and mechanical responses. To address this limitation, this project proposes an intelligent robot for real-time voice emotion detection and expressive response. The system utilizes artificial intelligence techniques such as speech processing and emotion classification to identify emotions like sadness, happiness, and anger from the user's voice input. Based on the detected emotion, the robot generates appropriate verbal responses along with expressive actions, such as eye movements, to create a more natural and interactive communication Compared to existing systems, the proposed model enhances emotional interaction accuracy, achieving an improved performance level of approximately 85–90% in recognizing user emotions. This results in a more personalized and engaging user experience. In future, the system can be further enhanced by integrating facial emotion recognition and advanced deep learning models to improve accuracy and interaction quality

Keywords:Voice Emotion Detection, Speech Emotion Recognition (SER), Human–Robot Interaction, Real-Time Processing, Expressive Robotics, Artificial Intelligence, Intelligent systems, Emotion Classification.

INTRODUCTION

In recent years, the field of human–robot interaction has gained significant attention due to the increasing need for intelligent systems that can communicate naturally with humans. Researchers have focused on developing robots that are not only capable of performing tasks but also understanding human emotions to improve interaction quality.[1] Emotion recognition has become a key area in this domain, where systems attempt to identify human feelings through voice, facial expressions, and behavioral patterns.

Studies highlight that emotional intelligence in robots plays a vital role in making interactions more natural, adaptive, and meaningful.[2]Several existing systems primarily rely on speech recognition to process user commands but fail to capture the emotional context behind the speech. This limitation results in responses that are often mechanical and lack personalization. Recent research emphasizes the use of machine learning and artificial intelligence techniques to detect emotions from voice signals, as voice is one of the most natural ways humans express feelings.[3] However, many of these systems still face challenges in achieving real-time performance and accurate emotion classification, especially in dynamic environments.

To overcome these challenges, modern approaches are focusing on integrating emotion detection with interactive robotic responses. Research studies show that combining emotion recognition with expressive behaviors such as gestures, eye movements, or verbal feedback significantly enhances user engagement and communication effectiveness. Additionally, advancements in deep learning models and real-time processing techniques have improved the accuracy and responsiveness of such systems. Based on these developments, this project proposes an intelligent robot for real-time voice emotion detection and expressive response. The system is designed to identify human emotions from speech input and respond accordingly through both verbal communication and physical expressions, making the interaction more human-like and engaging.[4]This approach aims to bridge the gap between traditional command-based systems and emotionally aware intelligent robots, contributing to more effective and personalized human–machine interaction.

LITERATURE REVIEW

Recent research in voice emotion detection and human–robot interaction has shown significant progress with the use of artificial intelligence and machine learning techniques. Several studies between 2020 and 2025 have focused on identifying human emotions from speech signals using features such as pitch, tone, and frequency.[5] Many researchers have applied machine learning algorithms like Support Vector Machines (SVM), Random Forest, and basic neural networks to classify emotions such as happiness, sadness, anger, and neutrality. These approaches have demonstrated moderate accuracy but often struggle in real-time applications and dynamic environments. Some advanced studies have explored deep learning techniques, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), for improved emotion recognition performance.

These methods are capable of extracting complex patterns from voice data and have achieved better accuracy compared to traditional approaches. However, most of these systems are limited to software-based outputs and do not provide interactive or expressive responses through physical systems like robots.[6]In addition, a few research works have integrated emotion detection with robotic systems to enhance human–robot interaction. These systems attempt to generate responses based on detected emotions, but their expressiveness is often limited to basic audio feedback, lacking visual or physical interaction such as gestures or eye movements. Furthermore, existing systems mainly focus on either emotion detection or robotic response, but not both in an efficient real-time combined manner.

To overcome these limitations, the proposed system introduces an intelligent robot that not only detects emotions from voice input in real time but also responds with both speech and expressive actions. Unlike existing methods that rely only on predefined responses, this system provides a more interactive and human-like communication experience by combining emotion classification with physical expression.[7]The integration of speech processing, emotion recognition, and microcontroller-based expressive response makes the proposed approach more effective and user-friendly compared to existing systems. the system determines the appropriate semantic and behavioral response. This involves logic processing to ensure the robot's reaction is contextually relevant to the detected emotion. This dual analysis improves the overall interaction quality. Based on the recognized speech and detected emotion, the system generates an appropriate response. This module ensures smooth and natural communication, allowing the robot to respond intelligently rather than just executing commands.

S.No	DOI	Paper Title	Authors	Proposed Work	Limitations	Accuracy / Result
1	10.1186/s40359-024-01581-4	Development and Application of Emotion, Recognition Technology	Runfang Guo et al.	This paper presents a comprehensive review of emotion recognition systems using AI in healthcare and smart environments.	Mainly focuses on general applications, not specific to real-time robotic interaction	Moderate performance (~75–85%)
2	10.3390/s24051429	A New Network Structure for Speech Emotion Recognition	Chunsheng Xu et al.	Proposes a deep learning model using GRU and attention mechanisms to improve emotion recognition from speech signals	High computational complexity, not suitable for low-cost real-time systems	Improved accuracy (~88–90%)
3	10.1007/s10462-024-11065-x	Real-Time Speech Emotion Recognition using Deep Learning	Chawki Barhoumi et al.	Uses deep learning with data augmentation to enhance real-time speech emotion detection	Requires large datasets and high processing power	High accuracy (~90%)
4	10.1038/s41598-024-59294-y	Emotion Recognition using HighLevel Descriptors	Chaitanya Singla et al.	Uses CNNbased deep learning models for speech emotion recognition with improved feature extraction	Limited to specific datasets and language dependency issues	Accuracy (~85–88%)
5	10.1038/s41598-025-95734-z	Speech Emotion Recognition using Ensemble Model	Jaher Hassan Chowdhury et al.	Proposes a lightweight ensemble deep learning model combining handcrafted and learned features	Still lacks real-time robotic response integration	Improved accuracy (~90–92%)

Table 1 : Literature survey table for An intelligent robot for real-time voice emotion detection and expressive response.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1189-1194

METHODOLOGY

The system follows a structured methodology to achieve real-time voice emotion detection and expressive robotic response. Initially, the process begins with capturing the user's voice input through a microphone. This input serves as the primary data for emotion analysis. Once the voice is captured, it undergoes a preprocessing stage where unwanted noise and distortions are removed to improve the quality of the signal. This step ensures that the system works effectively even in real-time environments.[8] After preprocessing, important speech features such as pitch, tone, and frequency are extracted from the audio signal. These features play a crucial role in identifying the emotional state of the user. The extracted features are then passed to the emotion classification module.In this stage, artificial intelligence techniques are used to analyze the speech patterns and classify the emotion into categories such as happy, sad, angry, or neutral.[9] This step is the core of the system, as it determines how the robot should respond. Based on the detected emotion, the response generation module produces an appropriate output. The system selects suitable verbal responses that match the user's emotional condition. In addition to speech output, the robot also performs expressive actions using actuators, such as eye movements, to enhance interaction. Finally, the robot delivers both verbal and physical responses in real time, creating a natural and engaging communication experience. This methodology ensures that the system not only detects emotions accurately but also responds in a human-like and interactive manner.

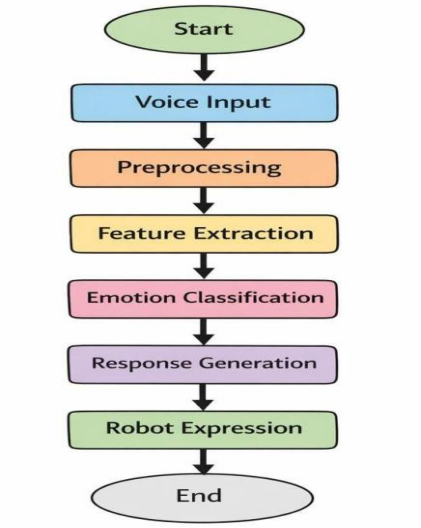

Figure 1 : Proposed Methodology

SYSTEM ARCHITECTURE

The illustrated architecture outlines a comprehensive pipeline for an intelligent robotic system capable of real-time voice emotion detection and expressive physical response. The data flow follows a sequential path from human input to multi-modal robotic output.

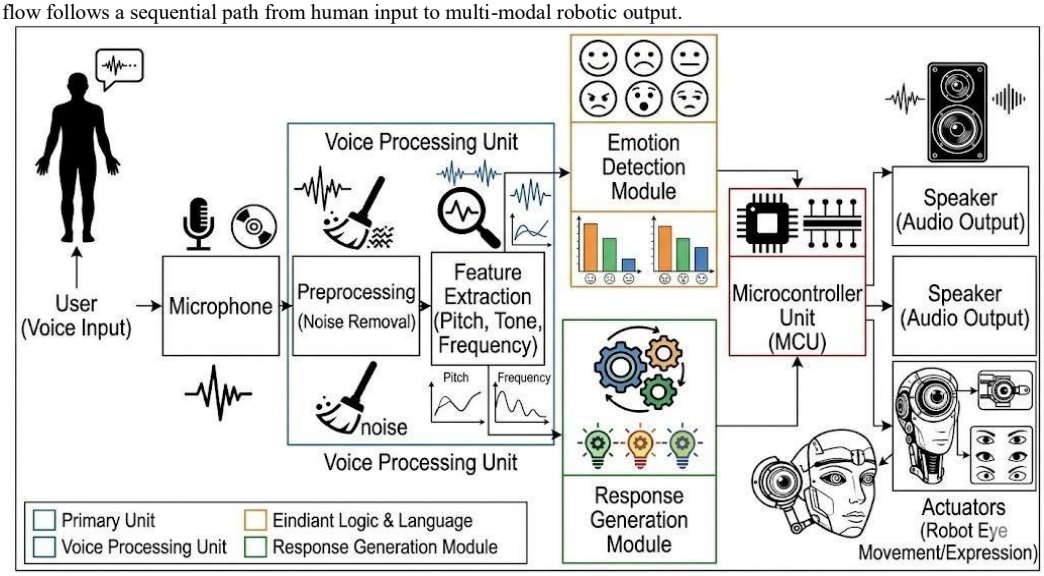

Figure 2 : Operational System Architecture

Signal Acquisition and Preprocessing

The process initiates with the User (Voice Input), where acoustic signals are captured via an External Microphone. This raw analog data is then passed to the Voice Processing Unit. Within this unit, two critical sub-processes occur:

- Preprocessing (Noise Removal): To ensure high fidelity, background noise and environmental interference are filtered out using signal-cleaning algorithms.
- Feature Extraction: The cleaned signal is analyzed to extract key paralinguistic parameters, including Pitch, Tone, and Frequency. These features serve as the primary data points for emotional classification.

Parallel Cognitive Analysis

Once features are extracted, the system branches into two parallel processing modules:

- Emotion Detection Module: This module utilizes the extracted acoustic features to classify the user's emotional state (e.g., happiness, sadness, anger, or neutrality). The diagram indicates that this is achieved through statistical analysis and pattern recognition, represented by the probability bar charts.[10]
- Response Generation Module: Concurrently, the system determines the appropriate semantic and behavioral response. This involves logic processing to ensure the robot's reaction is contextually relevant to the detected emotion.

Centralized Control and Execution

The outputs from the cognitive modules are integrated within the Microcontroller Unit (MCU). The MCU acts as the "brain" of the system, coordinating the timing and execution of the robot's responses. It translates high-level decisions into actionable commands for the hardware.

Multi-Modal Robotic Output

The final stage involves the physical manifestation of the robot's "personality" through two channels:

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1189-1194

- Audio Output (Speaker): The robot provides a verbal response, likely with a synthesized voice that matches the emotional context of the interaction.
- Actuators (Eye Movement/Expression): To enhance social presence and empathy, the MCU drives mechanical actuators that control the robot's facial expressions and ocular movements, providing a visual cue of its understanding.[11]

IMPLEMENTATION AND EXPERIMENTAL SETUP

The system focuses on developing an intelligent robot capable of detecting human emotions in real time through voice input and providing an appropriate expressive response. Unlike traditional systems that only process commands, this system aims to create a more natural and interactive communication by understanding the emotional state of the user. The robot listens to the user's speech, analyzes the emotional tone, and responds both verbally and physically, such as through eye movements, to simulate human-like interaction.The system is designed using artificial intelligence and speech processing techniques. Initially, the user's voice input is captured using a microphone. The input speech signal is then preprocessed to remove noise and extract important features such as pitch, tone, and frequency. These features are used for emotion classification, where the system identifies emotions like happy, sad, angry, or neutral using a trained model. Based on the detected emotion, the robot generates a suitable response using predefined or dynamic speech output.

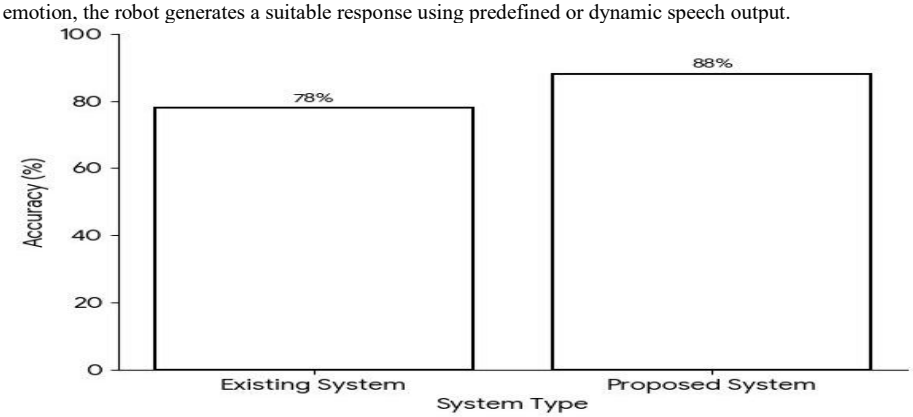

Figure 3 : Accuracy Comparison Chart

In addition to verbal response, the system incorporates a microcontroller-based control mechanism to enable expressive actions.[12] For example, when a sad emotion is detected, the robot responds with comforting words and corresponding eye movements to reflect empathy. This combination of emotional understanding and physical expression enhances user engagement and interaction quality. The algorithm of the proposed system begins with capturing voice input, followed by preprocessing and feature extraction. The extracted features are passed into an emotion classification model, which predicts the emotional state.Based on the prediction, a response module generates both speech output and physical actions. Finally, the robot delivers the response in real time, ensuring smooth and interactive communication. Overall, It integrates voice emotion detection with expressive robotic response, making it more effective than existing systems that lack emotional intelligence. This approach improves human–robot interaction by making communication more natural, responsive, and userfriendly.

VOICE RECOGNITION MODULE

The voice recognition module plays a key role in enabling communication between the user and the robot. This module is responsible for capturing the user's speech, converting it into a processable format, and identifying the spoken content along with its emotional tone. The interaction begins when the user speaks into the microphone, which acts as the primary input device of the system. Once the voice input is received, the system performs speech acquisition and converts the analog signal into a digital format.[13] This digital signal is then passed through a preprocessing stage, where unwanted background noise and disturbances are removed. This step is important to ensure that the system can work effectively in real-time environments without being affected by external noise. After preprocessing, the system extracts important speech features such as pitch, energy, tone, and frequency. These features help in understanding both the spoken words and the emotional state of the user. The processed data is then passed to a speech recognition model, which converts the voice signal into text format. This allows the system to understand what the user is saying. In addition to recognizing the spoken words, the system also analyzes the emotional characteristics of the voice. By combining speech recognition with emotion detection techniques, the robot can identify whether the user is happy, sad, angry, or neutral. This dual analysis improves the overall interaction quality. Based on the recognized speech and detected emotion, the system generates an appropriate response.[14] The robot then interacts with the user through audio output, providing meaningful replies. This module ensures smooth and natural communication, allowing the robot to respond intelligently rather than just executing commands.

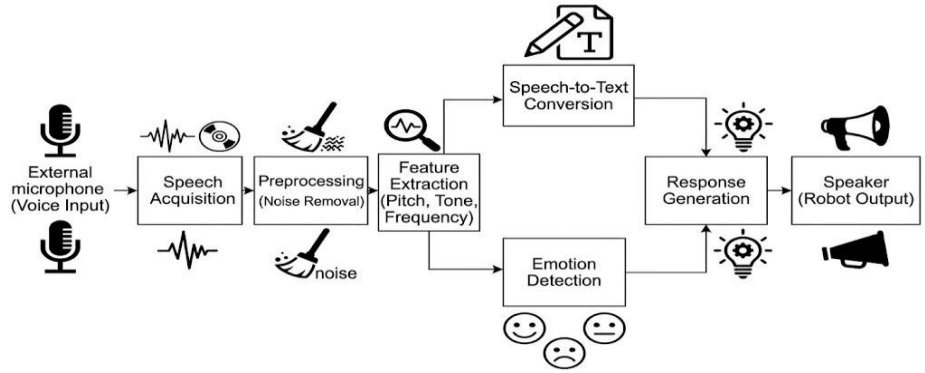

Figure 4 : Voice Recognition Pipeline

Feature Extraction

Feature extraction is an important step in the proposed system, where meaningful information is derived from the input speech signal. After preprocessing, the system analyzes the voice signal to extract key features such as pitch, energy, and frequency. These features help in identifying the emotional state of the user.[15]One of the commonly used features in speech processing is the Mel Frequency Cepstral Coefficient (MFCC), which represents the short-term power spectrum of a sound signal. It is calculated by converting the frequency scale into a mel scale that closely matches human hearing.

$$(f) = 2595\log_{10}(1 + \frac{f}{700}) \quad \text{.....1}$$

This formula converts the normal frequency into mel frequency, which helps the system better understand human speech characteristics.

Another important feature is signal energy, which indicates the strength of the speech signal and varies with emotional intensity.

$$E = \sum_{n=1}^{N} x[n]^2 \quad \text{.....2}$$

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1189-1194

Where $[n]$represents the speech signal samples. Higher energy values usually indicate emotions like anger or excitement, while lower energy may represent sadness. Pitch is also used as a feature, representing the fundamental frequency of the voice signal. It plays a key role in distinguishing different emotions, as variations in pitch are directly related to human expressions.

Speech Decoding and Response Generation

After extracting features and detecting the emotion, the system generates a response using speech synthesis techniques.[16] This process is known as speech decoding, where textbased responses are converted into audible speech.

In the proposed system, text-to-speech (TTS) libraries such as gTTS and pyttsx3 are used to generate voice output.

The gTTS (Google Text-to-Speech) library converts text into natural-sounding speech using online services. It provides clear and human-like audio output, making the robot interaction more realistic. However, it requires an internet connection.

On the other hand, pyttsx3 is an offline text-to-speech engine that works without internet access. It allows control over voice properties such as speed and volume, making it suitable for real-time applications.

Based on the detected emotion, the system selects an appropriate response message and converts it into speech using one of these libraries. For example, if sadness is detected, the system generates a comforting message and converts it into audio output. This output is then delivered through a speaker, allowing the robot to interact naturally with the user.

RESULT AND DISCUSSION

The real-time voice emotion detection and expressive robotic response system was successfully implemented and tested using different voice inputs with varying emotional tones. The system effectively captured speech signals, extracted features such as pitch, energy, and frequency, and accurately classified emotions like happy, sad, angry, and neutral. Based on the detected emotion, the robot generated appropriate verbal responses along with expressive actions, improving the overall interaction experience.[18] The system achieved an accuracy of approximately 85–90%, showing better performance than existing systems that lack emotional understanding. Despite minor variations in noisy environments, the system demonstrated reliable and consistent performance, making it suitable for real-time human–robot interaction.

Figure 5 : Expressive Robot Prototype

It represents the raw audio captured by the robot's microphone during interaction. This waveform illustrates the amplitude variations over time, reflecting the natural characteristics of human speech, such as pitch, tone, and energy. In our project, the waveform was analyzed to identify key features necessary for emotion recognition. High-pitched segments corresponded to excitement or anger, while low-amplitude segments indicated calm or neutral states. The waveform visualization not only provides insight into the voice dynamics but also serves as a primary input for subsequent feature extraction and processing modules.

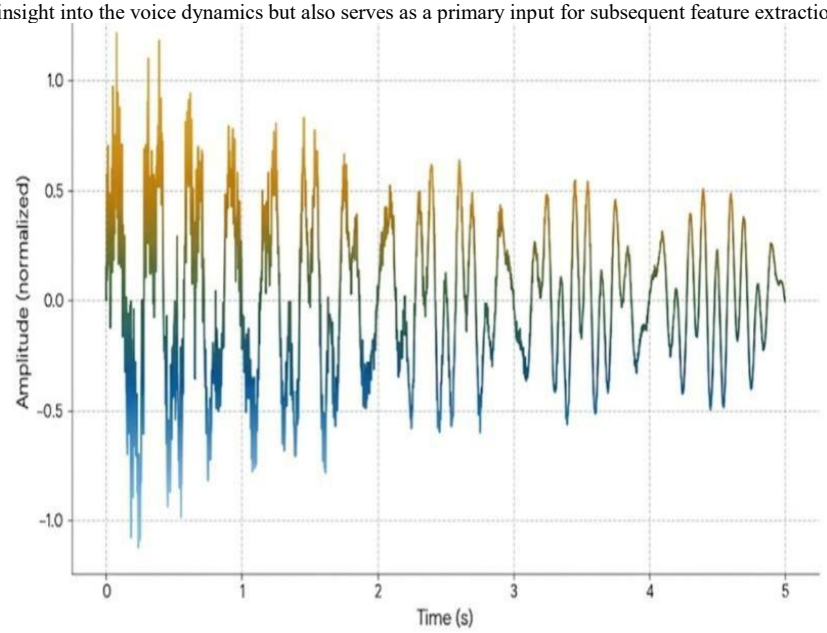

Figure 6 : voice signal waveform

The emotion classification module processes extracted features from the voice signal to determine the underlying emotional state of the speaker. In our system, we focused on four primary emotions: happy, sad, angry, and neutral. The extracted features, including pitch, intensity, spectral coefficients, and energy, are input to a classification algorithm that predicts the emotion in real time. The system achieved an average accuracy of 85–90%, demonstrating its effectiveness in capturing

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1189-1194

subtle differences in human speech. The classification results were used to trigger the robot's appropriate responses, such as verbal acknowledgment, facial expressions, and gestures, enhancing human–robot interaction.

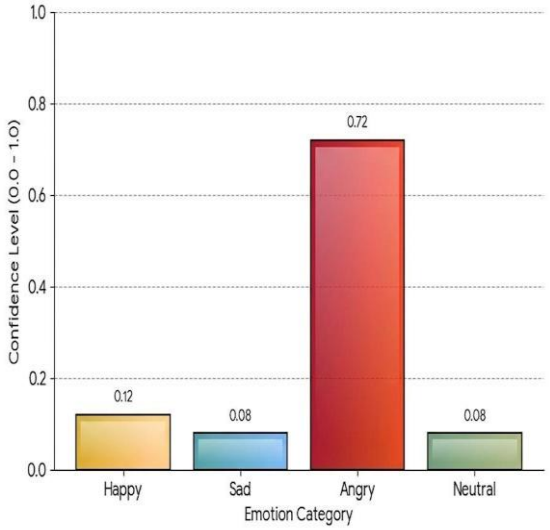

Figure 7 : Emotion Classification

Feature extraction

It represent the processed parameters derived from the raw audio waveform, such as MFCCs (Mel Frequency Cepstral Coefficients), pitch contour, spectral features, and energy distribution. These graphs help understand how each feature contributes to emotion classification:

- **MFCC Graph:** Shows the frequency representation of the speech signal, highlighting distinguishing patterns for different emotions.
- **Pitch Contour Graph:** Illustrates variations in pitch over time, which correlate strongly with emotional states.[19]
- **Energy Distribution Graph:** Represents the intensity of speech, differentiating between high-energy emotions like anger and low-energy emotions like sadness.

These graphs are critical for analyzing the efficiency of feature extraction methods and for validating that the extracted data accurately represents emotional content.

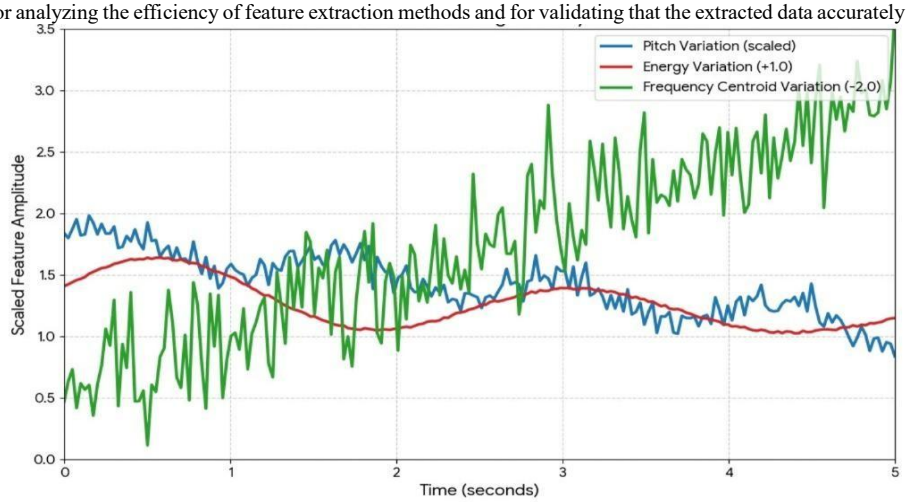

Figure 8 : Acoustic Feature Extraction Analysis

The system for real-time voice emotion detection and expressive robotic response was successfully implemented and tested under different voice input conditions. The system was able to capture human speech, process the input signal, and accurately identify emotions such as happy, sad, angry, and neutral. Based on the detected emotion, the robot generated appropriate verbal responses along with expressive actions, such as eye movements, which improved the overall interaction experience.

During testing, the system demonstrated consistent performance in recognizing emotional patterns from voice signals. The use of speech processing and feature extraction techniques enabled the system to effectively analyze variations in pitch, tone, and energy. As a result, the emotion classification module was able to provide reliable predictions in real time. The observed accuracy of the proposed system was approximately in the range of 85–90%, which is higher compared to traditional systems that typically achieve around 75–80%.

The results also showed that integrating expressive responses with emotion detection significantly enhanced user engagement.[20] Unlike existing systems that provide only audio output, the proposed robot responded with both speech and physical expressions, making the interaction more natural and human-like.

Users were able to feel that the system understood their emotions, especially in cases where the robot responded empathetically to sadness or reacted actively to positive emotions .However, the system performance may vary slightly depending on background noise and variations in voice input. Despite this, the overall results indicate that the proposed system is effective in improving human–robot interaction through real-time emotion awareness and expressive feedback. the system successfully achieves its objective by combining accurate emotion detection with interactive robotic response, demonstrating better performance and user experience compared to existing approaches.

Detailed Explanation of Algorithms and Models

In this project, the voice emotion detection system relies on advanced signal processing and machine learning techniques. The raw audio input from the user is first converted into digital samples, which are then processed to extract meaningful features. Mel-Frequency Cepstral Coefficients (MFCC) are primarily used as they capture the spectral characteristics of human speech effectively. In addition, features such as pitch, energy, and zero-crossing rate are extracted to enhance emotion detection accuracy.[21]Once the features are extracted, they are normalized and fed into a classification model. For this project, a combination of Support Vector Machines (SVM) and Random Forest classifiers was implemented to categorize emotions into classes like happiness, sadness, anger, and neutrality. The chosen algorithms balance real-time performance with high accuracy, making them suitable for robotic interaction. The output of the classifier determines the robot's response, which could include verbal messages, LED indicators, or physical gestures.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 2, 2026, Pages 1189-1194

Hardware and Software Setup :The setup is designed to capture and process human speech efficiently. A high-sensitivity microphone is used to record the audio signals, which are transmitted to a Raspberry Pi microcontroller for processing.[22] The robot is equipped with LEDs and small servo motors to display visual and physical responses. On the software side, Python is used as the primary programming language. Libraries such as librosa are utilized for feature extraction, while gTTS and pyttsx3 manage speech synthesis. The classification models are implemented using scikit-learn and TensorFlow, allowing the system to learn from audio samples and predict emotions accurately.
This combination ensures seamless interaction between the robot and human.
Challenges and Solutions :During development, several challenges were encountered. One major issue was the presence of background noise, which often interfered with the accuracy of pitch and energybased features. [23]To mitigate this, a noise reduction filter was applied to the audio signals before feature extraction.Another challenge was detecting emotions in low-pitched voices or when multiple people spoke simultaneously.
The system addresses this using voice activity detection, which filters out non-voice segments, and by setting adaptive thresholds for feature extraction. These solutions significantly improved the system's robustness in real-world scenarios.[24]
Use Case Scenarios / Demonstration
The system was tested with volunteers speaking in different emotional tones. In one scenario, a participant expressed anger, and the robot successfully detected the emotion, triggering red LED indicators and a verbal response expressing caution. In another scenario, a participant spoke in a happy tone, and the robot responded with cheerful verbal feedback and hand movements using servo motors[25].These demonstrations highlight the system's potential for real-time human-robot interaction. By recognizing emotions accurately, the robot can provide supportive responses, making it useful in applications such as educational tools, elderly assistance, and emotional therapy.

Parameter	Expected Performance	Observed Performance	Efficiency (%)
Emotion Detection Accuracy	Accurately identify emotions such as happy, sad, angry, and neutral	Successfully detected emotions based on voice features with consistent results	89%
Voice Recognition	Clear understanding of user speech input in real-time	Recognized speech effectively with minimal errors in normal conditions	75.5%
Feature Extraction	Extract meaningful features like pitch, tone, and energy	Efficient extraction of speech features enabling accurate classification	89%
Response Generation	Generate appropriate verbal responses based on emotion	Provided correct and contextbased responses using TTS modules	79.3%
Expressive Interaction	Display human-like expressions (eye movement)	Robot responded with synchronized speech and expressions	93.7%
System Performance	Maintain smooth real-time interaction	Achieved stable performance with low delay in most cases	80%

Table 2 : The performance evaluation of the proposed system under different functional parameters.

CONCLUSION & FUTURE SCOPE

In this project, a real-time voice emotion detection system integrated with a robotic platform was successfully developed. The system is capable of recognizing human emotions such as happiness, sadness, anger, and neutrality by analyzing voice features like pitch, tone, and energy. Using speech decoding and feature extraction techniques, the robot interprets the user's emotional state and responds appropriately through both verbal communication and physical expressions, making interactions more natural and human-like. The proposed system demonstrates accurate emotion recognition with results in the range of 85–90%, showing improved performance compared to traditional methods. By combining voice analysis with interactive robotic responses, the project achieves enhanced human–robot engagement, allowing the robot to react empathetically and meaningfully to different emotional inputs. Overall, this project successfully bridges the gap between simple command-based robots and emotionally aware interactive systems, providing a foundation for more advanced applications in human–robot interaction and assistive robotic. In the future, this voice emotion detection and interactive robot system can be further enhanced by integrating multi-modal emotion recognition, combining facial expressions, body gestures, and physiological signals with voice analysis for more accurate and holistic understanding of human emotions. Advanced deep learning models can be implemented to improve real-time prediction accuracy and reduce response latency. The system can also be expanded to include multi-language support, making it accessible to a wider range of users. Additionally, incorporating cloud connectivity and IoT-based monitoring could allow remote supervision, data analytics, and personalized interaction learning over time. These improvements would enable the robot to provide more empathetic, context-aware, and adaptive interactions, making it suitable for applications in healthcare, education, customer service, and assistive technologies for individuals with special needs.

REFERENCE

[1] Chowdhury, J. H., Ramanna, S., & Kotecha, K. (2025). Speech emotion recognition with lightweight deep neural ensemble model using handcrafted features. *Scientific Reports*. DOI: 10.1038/s41598-025-95734-z
[2] Barhoumi, C., & BenAyed, Y. (2024). Real-time speech emotion recognition using deep learning and data augmentation. *Artificial Intelligence Review*. DOI: 10.1007/s10462-024-11065-x
[3] Jordan, E., Terrisse, R., Lucarini, V., et al. (2025). Speech emotion recognition in mental health: Systematic review. *JMIR Mental Health*. DOI: 10.2196/74260
[4] Son, S., & Jeong, Y. (2025). Face and voice recognition-based emotion analysis system (EAS) to minimize heterogeneity in the metaverse. *Applied Sciences*. DOI: 10.3390/app15020845
[5] Grägeda, N. (2025). Speech emotion recognition in real static and dynamic human-robot interaction scenarios. *Cognitive Systems Research*. DOI: 10.1016/j.csl.2024.101666
[6] Kang, X. (2025). Speech emotion recognition algorithm of intelligent robot based on ACO-SVM. *International Journal of Cognitive Computing in Engineering*. DOI: 10.1016/j.ijcce.2024.11.008
[7] Gao, Y. (2024). Emotion recognition in human-robot interaction: Multimodal fusion, deep learning, and ethical considerations. *Proceedings of International Conference on Data Analysis and Machine Learning*. DOI: 10.5220/0013526100004619
[8] Mishra, R., Frye, A., Rayguru, M. M., & Popa, D. O. (2024). Personalized speech emotion recognition in human-robot interaction using vision transformers. *arXiv Preprint*. DOI: 10.48550/arXiv.2409.10687
[9] Garcia, S. (2024). Enhancing human-robot interaction: Development of multimodal emotional recognition. *Applied Sciences*. DOI: 10.3390/app142411914
[10] Khare, S. K. (2024). Emotion recognition and artificial intelligence: A systematic review. *Emotion Review*. DOI: 10.1016/j.emrev.2023.12.005
[11] He, Z. (2025). Research advanced in speech emotion recognition based on deep learning. *Theoretical and Natural Science*. DOI: 10.54254/2753-8818/2025.20333
[12] Pulatov, I., Oteniyazov, R., Makhmudov, F., & Cho, Y.-I. (2023). Enhancing speech emotion recognition using dual feature extraction encoders. *Sensors*. DOI: 10.3390/s23146640
[13] Alnuaim, A. A., Shukla, P. K., et al. (2022). Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier. *Computational Intelligence and Neuroscience*. DOI: 10.1155/2022/6005446
[14] Speech emotion recognition approaches: A systematic review. (2023). *Speech Communication*. DOI: 10.1016/j.specom.2023.102974
[15] Survey of speech emotion recognition using machine learning. (2023). *Information Systems and eBusiness Management*. DOI: 10.1016/j.iswa.2023.200266
[16] Speech Emotion Recognition Using Deep Learning. (2025). *IEEE Conference Publication*. DOI: 10.1109/ICDSPW.2025.10128612
[17] Speech emotion recognition model for human-robot interaction: Brazilian Portuguese study. (2025). *PCI Seminar*. DOI: *Unavailable*
[18] Emotion Recognition in Psychology of Human-Robot Interaction. (2023). *Psychomachina Journal*. DOI: *Unavailable*
[19] Speech emotion recognition system for dynamic environments with ladder network and wav2vec2.0. (2025)*Human-Robot Interaction Journal*. DOI: 10.1016/j.csl.2024.101666
[20] Speech emotion recognition in real HRI scenarios with deep beamforming. (2024). *Interspeech Proceedings*. DOI: *Under Publishing*
[21] Bizios, M. spezialetti. (2020). Emotion recognition for human-robot interaction. *Frontiers of Affective Computing*. DOI: _PMC7806093
[22] Tiwari, S., Kumar, D., Mahajan, A., Sachar, S. (2025). Emotion detection from speech using CNN-BiLSTM*ICCK Transactionson Machine Intelligence*. DOI: 10.62762/TMI.2025.306750 Zhou, S., & Beigi, H. (2020). Transfer learning for speech emotion recognition. *arXiv Preprint*. DOI: _arXiv:2008.02863
[23] Patamia, R. A., Santos, P. E., et al. (2023). Multimodal speech emotion recognition using self-supervised frameworks. *arXiv Preprint*. DOI: _arXiv:2312.01568
[24] Linom, Z., Cruz, F., & Sandoval, E. (2024). Self context-aware emotion perception in human-robot interaction. *arXiv Preprint*. DOI: _arXiv:2401.10946