Robotica (2025), **43**, pp. 2143–2171
doi:10.1017/S0263574725000438

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Voice-enabled human-robot interaction: adaptive self-learning systems for enhanced collaboration

Indra Kishor[1], Udit Mamodiya[2], Sumit Saini[3] and Badre Bossoufi[4]

[1]Department of CSE, Poornima Institute of Engineering & Technology, Jaipur, Rajasthan, India
[2]Faculty of Engineering & Technology, Poornima University, Jaipur, Rajasthan, India
[3]Department of Electrical Engineering, Central University of Haryana, Mahendergarh, Haryana, India
[4]LIMAS Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohammed Ben Abdellah University, Fez, Morocco
Corresponding author: Sumit Saini; Email: drsumiteed@cuh.ac.in

Received: 30 January 2025; **Accepted:** 27 February 2025; **First published online:** 23 May 2025

Keywords: human-robot interaction; voice recognition; self-learning; emotional intelligence; reinforcement learning

Abstract

This research proposes an adaptive human-robot interaction (HRI) that combines voice recognition, emotional context detection, decision-making, and self-learning. The aim is to overcome challenges in dynamic and noisy environments while achieving real-time and scalable performance. The architecture is based on a three-stage HRI system: voice input acquisition, feature extraction, and adaptive decision-making. For voice recognition, modern pre-processing techniques and mel-frequency cepstral coefficients are used to robustly implement the commands. Emotional context detection is governed by neural network classification on pitch, energy, and jitter features. Decision-making uses reinforcement learning where actions are taken and then the user is prompted to provide feedback that serves as a basis for re-evaluation. Iterative self-learning mechanisms are included, thereby increasing the adaptability as stored patterns and policies are updated dynamically. The experimental results show substantial improvements in recognition accuracy along with task success rates and emotional detection. The proposed system achieved 95% accuracy and a task success rate of 96%, even against challenging noise conditions. It is apparent that emotional detection achieves a high F1-score of 92%. Real-world validation showed the system's ability to dynamically adapt, thus mitigating 15% latency through self-learning. The proposed system has potential applications in assistive robotics, interactive learning systems, and smart environments, addressing scalability and adaptability for real-world deployment. Novel contributions to adaptive HRI arise from the integration of voice recognition, emotional context detection, and self-learning mechanisms. The findings act as a bridge between the theoretical advancements and the practical utility of further system improvements in human-robot collaboration.

1. Introduction

The field of robotics has experienced explosive growth over the last few decades, from mechanical tasks to very complex, adaptive systems that could interact with people. As robots become present in healthcare, education, and industrial work, the prospects for easier and more effective interaction between humans and robots grow immensely. While traditional automation is about designing a system to perform predefined functions, modern robotics is about collaboration with humans and requires adaptability, emotional intelligence, and seamless communication. However, considerable progress notwithstanding, working toward these goals still presents a huge challenge, especially in dynamic and unstructured environments. The interaction through voice, a completely natural and intuitive medium, is one of the key facets of human-robot interaction (HRI). Because of this, voice-activated robots are far more accessible and simplified for the user, as humans mostly voice their feelings. Variability of human-involved speech, owing to accents, tonal shifts, and interference from background noise, presents major technical challenges to current voice recognition systems. Most automatic speech recognition

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

systems currently in use have been constructed on large and standard-based datasets, typically serving as a pre-training reference, essentially limiting the adaptation of the many different users in their various scenarios. For instance, a system that is trained under standard American English might have great problems in recognizing and interpreting non-native-sounding accents or dialects. Furthermore, during and after deployment, these systems do not possess self-improving features to evolve in their run-time stage that satisfy the expectation of delivering an appropriate performance in real-life settings. Emotional intelligence remains another crucially important yet weakly covered area of HRI. A large part of human communication depends on emotional cues, adding tone, context, and value to speech. For example, a respondent to an agitated user could calm him while a tone of elation could similarly respond. While emotional recognition has been extensively explored in natural language processing (NLP) and human-computer interaction, there is still little appreciation for levels of intelligence integration in robotics. Most existing robots lack contextual knowledge of the emotional state of their users, such that interactions may appear mechanical or depriving of personal experience. AI and ML have sparked the development of new opportunities to tackle some of those issues. Developments in deep learning models, especially those based on transformer architectures, have transformed the landscape of voice recognition. For instance, Wav2Vec and Deep Speech have made great strides in performance by handling complex acoustic models and feature extraction. Similarly, advanced emotional recognition based on NLP has been achieved in speech pattern and tone recognition along with semantics inference of user emotional states. While impressive for each by itself, in concert with developing adaptive HRI, a lot still stands to be differential. Particularly interesting is leveraging the ability of a system to learn autonomously. Self-learning systems usually powered by reinforcement learning (RL) allow robots to learn over time and improve their performance with interactions and feedback. Unlike static models, self-learning systems dynamically adapt to new users and environments without manual updates or retraining. For instance, a robot may learn a user's unique accent after a few interactions or finetune its responses based on corrective feedback. This adaptability lends enhanced usability to the robot, maintaining performance through diverse and dynamic conditions. Integrating self-learning, speech recognition, and emotional intelligence into a concise HRI paradigm establishes a wholly unified framework for the context of these three that ultimately translates to an interaction centered "naturally" much around human-style interactions. Adaptive systems will have to learn in every instance from the changes in user preferences, the kind of context found in an interaction, and whether or not doing so enhances task completion. This capability is very much prominent in fields such as healthcare, in which robots have been increasingly positioned as they deploy themselves to monitor patients, provide therapy, and accompany a companion. For instance, a robot assisting an adult citizen could modulate its voice and communication style as per the emotional state of the patient, hence providing comfort and reduced stress. Likewise, in educational terrain, robots would shift their responses to attune themselves to the students' emotional states, thus inducing a supporting learning environment. In spite of these anticipated advancements, a number of challenges have stood between them and their users. A robust voice recognition system must draw large amounts of computation power and fine-tuning to operate efficiently in a variety of noisy environments. Emotional intelligence has many draw factors beyond being theoretically attractive; however, processing data on emotions is probably the hardest to crack, thus meeting severe ethical questions when interpreting delicate user emotions. Self-learning systems are great but really have to take care to strike a balance between adaptability and stability, so they do not become erratic or undesired in their behavior. Integrating advanced robotics into daily life emphasizes adaptability in HRI. Voice-activated systems constitute an important part of HRI since they determine an interaction method that is natural and can be conducted without hands. These systems must then be extended beyond just recognizing basic commands, with dynamic adaptation to human behavior for seamless interaction. In this context, the problems are twofold: to enable robots to understand and respond appropriately to various commands given by users, as well as enable them to detect and respond appropriately to the emotional states of the users. Most current state-of-the-art systems mainly stress static interaction, where a robot performs a pre-programed set of tasks according to voice inputs. These systems perform quite effectively under controlled conditions. However, once environmental noise, speaker variability, and unpredictable

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

user behavior are introduced, the systems fall apart. For example, robots deployed in public environments can hardly process commands due to noise interference, while systems in personal environments throw in the towel whenever an atypical accent or modeling of speech is presented. This inability to learn in a dynamic fashion and learn beyond is what makes these solutions inadequate for complex real-life scenarios. The limitations placed by the traditional HRI systems are equally technical and user-center. Interactions devoid of emotional sensitivity tend to be stilted and impersonal, thereby undermining user satisfaction and acceptance. Emotionally intelligent robotics would thus describe the capability of a robot to perceive, understand, and react accordingly in a given interaction. This might mean, for example, being able to pick up frustration from the user's tone of voice and to alter the responses according to it; hence, a big deal about enhancing an interaction's effectiveness. Emotional intelligence is still a little explored dimension in the field of robotics because of its computational complexity and the lack of a standardized framework for emotion recognition and the generation of appropriate responses. Moreover, since robots are working in more dynamic, multi-user environments, self-learning becomes an important factor. Self-learning robots learn and then adapt their reception of voices and emotional cues to their varied users over time. Such systems exploit feedback from the user interaction to improve their performance in a more effective and versatile manner. For example, a self-learning robot deployed within a healthcare setting might alter its communication style according to the comfort level of a patient to provide a more personalized and empathetic experience. Robots can further enhance work task efficiency through learning of team dynamics and individual preferences for collaborative working environments. The theoretical basis of self-learning in robotics is derived from machine learning, that is, from the area known as reinforcement learning, or RL. RL is where the agent learns how to take appropriate actions in the environment to maximize rewards that in RL are regarded as rewards minus penalties. The application of HRI provides a means of enabling robots to learn from their interaction with humans and improve on their own end. Also, unlike models that are pre-trained and thus need to be retrained on the fly for new users or environments, and tasks, they are well-adapted to real-world settings. However, the integration of RL with voice recognition and emotional intelligence brings about challenges, such as balancing exploration and exploitation during learning, ensuring computational efficiency, and maintaining ethical considerations. Voice recognition and emotional intelligence, though individually significant, achieve their full potential when integrated into a unified HRI framework. This is where the synergy of these technologies helps robots to understand not only the content of the user command but also the context and the emotional undertone. For instance, a voice command like "I need help" may imply something different in the tone of the user, the urgency of the situation, and the emotional state of the user. For example, in eldercare, emotionally intelligent robots can keep people company, reduce loneliness, and enhance the mental well-being of older individuals. In learning settings, adaptive robots can shape their interactions in ways that would suit the learning needs of students, making education more inclusive and effective. Again, this gives rise to all-important questions around privacy, security of data, and ethical deployment of AI systems. The solution to these concerns will be of critical importance in the responsible deployment of advanced HRI systems.

This paper proposes a new framework for voice-enabled HRI to address the aforementioned challenges. The proposed framework integrates state-of-the-art voice recognition models with mood detection and RL to create an adaptive, emotionally intelligent robotic system. The validation of the proposed system based on synthetic datasets and simulation environments is performed using multiple metrics, such as recognition accuracy, reliability of mood detection, and completion rates of tasks. This research not only contributes to the development of more intuitive HRI systems but also sets the foundation for future advancements in human-centric robotics.

2. Literature review

HRI has recently exploded, fueled by important developments in enabling technologies: voice recognition, emotional intelligence systems, and mechanisms for self-learning. Design for HRI depends on metrics in performance that go as far as accuracy, adaptability, and user satisfaction within economic

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

feasibility as the base criteria for further integration. Voice recognition is central in enabling intuitive interaction between humans and robots. These are transformer-based architectures, such as Wav2Vec 2.0 [1] and Whisper [2], which improve recognition accuracy quite significantly, particularly in multilingual and noisy environments. Wav2Vec 2.0 achieves self-supervised learning so that it can generalize across varying acoustic conditions while Whisper emphasizes the robustness for transcribing low-resource languages [2]. Though the voice recognition system has undergone advancements, overlapping speech, and incomplete commands remain major obstacles. One work looked into the application of semantic analysis to a voice recognition processing pipeline to result in 10% better multi-speaker recognition accuracy. A critical limitation still lies in real-time deployment: the computational burden. Current works focus on using lightweight models: adaptations of MobileNet for speech tasks that offer the right trade-off between efficiency and accuracy [3]. Robots with emotional intelligence can connect better with their users by better interpreting and responding to emotional signals. Models such as RoBERTa and XLNet have been fine-tuned for sentiment and emotion analysis. These models yield very high accuracies in controlled environments [4, 5]. The author proposed a multimodal emotion recognition system that fused acoustic and linguistic cues, attaining an F1 score of 0.89 on the diverse datasets [6]. One of the main current limitations of emotion detection systems is their dependence on categorical labels (e.g., happy, sad, angry), thus failing to capture more nuanced or blended emotions in a more precise way. The author highlighted such a gap with his dimensional approach that maps emotions onto continuous spectra like arousal and valence [7]. The future research should focus on integration of physiological signals, like galvanic skin response and EEG data, to enhance robustness and interpretability in real-world interactions [8]. Self-learning capabilities are necessary for robots operating in dynamic and unpredictable environments. RL methods, such as Soft Actor-Critic and Proximal Policy Optimization (PPO), enable adaptive behavior through feedback-based learning. The author demonstrated that RL-based HRI systems adapted the user-specific preferences in five interactions and improved the task efficiency by 15%. However, the application of self-learning in multi-user and multi-task scenarios remains underexplored. Research by author extended RL frameworks to collaborative robots, enabling them to optimize task allocations dynamically [9]. While promising, these systems are computationally intensive and require significant training time. Meta-RL, which leverages prior experience to accelerate adaptation, holds the promise of breaking these bottlenecks [10]. Evaluation of voice recognition and emotion detection systems is often conducted using metrics like accuracy, precision, recall, and F1 scores. Accuracy improvements of up to 20% were reported when user feedback was incorporated into adaptive voice models [11]. In the absence of standardized evaluation protocols, cross-system comparisons are restricted. Reliability, often measured as task completion rates and robustness against environmental variability, is critical for HRI systems. Another study emphasized the need for context-aware reliability benchmarks to account for diverse operational conditions [12]. The reliability of emotional detection systems is assessed using multimodal datasets that include speech, text, and physiological signals [13]. The study introduced a benchmark dataset for emotion recognition in cross-cultural scenarios, highlighting the disparities in model performance across demographic groups. The lack of universally accepted benchmarks for evaluating emotional intelligence in HRI systems remains a significant research gap [14]. Adaptability, measured by the system's learning rate and generalization capability, is a crucial metric for self-learning systems. Metrics such as the number of iterations required for convergence and the rate of error reduction over time are commonly used [15]. While these metrics provide insight into system efficiency, they often neglect the trade-offs between adaptability and computational costs. Future research should establish scalability benchmarks that account for both learning efficiency and resource utilization. Economic feasibility depends on the affordability of hardware, computational requirements, and scalability of solutions. Cloud-based architectures, such as those reported in study [16]. Another study reduces upfront costs by offloading computational tasks to remote servers [17]. However, reliance on cloud infrastructure introduces latency and raises privacy concerns, particularly in sensitive domains like healthcare. Hybrid architectures combining edge and cloud computing have shown potential to address these challenges [18]. Another work demonstrated that hybrid systems reduced latency by 25% while maintaining cost-effectiveness, making them suitable for real-time applications in resource-constrained environments.

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

Resource-constrained environments, such as rural healthcare facilities, pose unique challenges for HRI systems [19]. Open-source hardware platforms like NVIDIA Jetson Nano and Raspberry Pi offer affordable alternatives, but their limited computational capacity restricts their applicability to basic tasks [20]. Efforts by author to optimize lightweight algorithms for such platforms have demonstrated promise but require further validation in real-world settings. As robots are increasingly deployed in dynamic and interactive settings, multimodal systems that combine various input modalities – such as speech, gestures, and facial expressions – have gained prominence [21]. These systems reinforce the robustness of communicated interactions and ultimately the illusiveness of these interactions under layers of complementary data sources. For example, the authors designed a multimodal framework that is capable of recognizing speech and gestures, finding over a 15% improvement in user's happiness scores as compared to those in unimodal architecture [22]. What they allow is robots to better execute a user command of greater complexity, especially in noisy contexts or where a user does not complete the verbal input. The integration of multimodal inputs has brought challenges with regard to computation, particularly in real-time contexts. Therefore, synchronizing data and maintaining low latency are two key areas for future work to address. Adaptive fusion techniques, allowing input modalities to be prioritized according to their context, have promising solutions to these issues [22]. There are serious challenges of scalability and generalization across heterogeneous settings. Applications requiring high trust and transparency such as healthcare or education have seen a rise in the implementation of explanations in AI (XAI) into HRI systems [23]. XAI aims to explain how AI models make decisions, thus developing trust and supporting debugging. An author proposed an XAI-supported emotion detection system that gives visual references for mood classification, thus improving user trust in its emotional intelligence capabilities [24, 25]. However, the integration of XAI in HRI is still in its infancy. Existing approaches primarily focus on post-hoc explanations, which provide insights after the decision-making process. There is a need for inherently interpretable models that provide real-time explanations during interactions. Additionally, balancing interpretability with performance remains a challenge, as simpler models often compromise accuracy [26]. Collaborative robots (cobots) are designed to work alongside humans, making adaptability and safety paramount. Recent advancements in adaptive control systems have enabled cobots to dynamically adjust their behavior based on user inputs and environmental changes. For example, a study demonstrated an RL-based control system that allowed cobots to adapt their motion trajectories in real-time, reducing collision rates by 25%. Safety, a critical factor in collaborative settings, has been addressed through the integration of real-time monitoring systems [27]. The author proposed a sensor fusion approach combining vision and proximity sensors to enhance collision avoidance [28]. Despite these improvements, achieving seamless human-robot collaboration requires further exploration of shared intent modeling, where robots anticipate user actions to proactively assist in tasks [29, 30]. The scalability of HRI systems remains a major challenge, particularly in terms of adapting to diverse users, environments, and tasks. Current systems often rely on high-performance hardware and computationally intensive models, making them cost-prohibitive for large-scale deployments [30]. Efforts by author [31] to optimize lightweight models for low-power devices have shown promise but require further validation in real-world applications. Additionally, the scalability of self-learning systems is hindered by the computational resources needed for RL. Distributed learning techniques, which leverage cloud and edge computing, have been proposed to address these challenges [32]. However, ensuring data security and minimizing latency in distributed architectures are critical areas for future research. HRI systems deployed in multilingual or cross-cultural settings often face challenges in adapting to diverse user preferences and communication styles. For instance, emotion recognition systems trained on Western datasets may fail to accurately interpret emotional expressions in Asian or African contexts. Research by [33] highlights the need for culturally diverse datasets to improve system generalizability. The focus of future research should be to develop adaptive learning frameworks that could fine-tune models dynamically in light of regional and cultural nuances. Academia-industry partnerships may be required to develop vast, multilingual datasets for training and evaluation purposes. Large-scale HRI deployments are accompanied by substantial ethical considerations about data privacy, algorithmic transparency, and bias mitigation. Voice and emotion recognition systems often rely on sensitive user

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

data, which, if mishandled, could lead to privacy violations. Developing secure data handling practices, such as encryption and anonymization, is essential for building user trust [34]. Bias in AI algorithms has also been an important weakness in general owing to the systems cannot learn from user interactions. The automated outcomes, based on biased training data, could be harming the fairness and progressive aspects of any HRI system. The methods of explainable AIs that have been proposed [35] can lessen the fear of discrimination towards explaining ability transparency and interpretability of the decisions made by AI.

2.1. Research gaps

Despite significant progress in HRI systems, some research gaps persist along with related enabling technologies, performance evaluation, and deployment in real-world applications. The establishment of these research gaps is critical for developing more robust, adaptive, and human-centric HRI systems.

2.1.1. Integration of technologies

One of the most significant chasms for HRI remains the integration among enabling technologies of voice recognition, emotional intelligence, and self-learning mechanisms. Where the advancements were quite significant at the individual levels, unified frameworks that combine capabilities seamlessly are indeed rare. For example, most voice recognition systems concentrate on speech accuracy but lack the emotional context and self-learning features to adjust according to user-specific accents or an evolving environment. This fragmented approach prevents the system from delivering integrated and intuitive interactions [36].

2.1.2. Adaptability and scalability

Humanoid robot-human interactivity systems have poor adaptability, especially in adaptable environments. They usually are trained on static datasets and are unable to generalize across various users who differ in attributes such as race, gender, socioeconomics, culture, and language. For example, a model that acts to recognize emotions could often be biased towards what was on a training dataset thereby culminating in poor performance within underrepresented groups, further compounding this issue. Further problems exist with respect to scalability. Many modern intelligent systems up to the level of deep learning-vocal recognition and RL may not be deployable in resource-constrained settings such as rural healthcare systems and low-cost educational environments because the computing loads are overwhelming [37].

2.1.3. Real-world benchmarking

The bulk of HRI research is performed in controlled laboratory settings, far removed from real-world challenges. Consequently, systems often fail to maintain performance under noisy environments, incomplete user inputs, or multitasking scenarios. Moreover, a lack of standard benchmarks for testing adaptability, emotional detection accuracy, and task efficiency complicates the assessment and comparison of HRI systems across studies [38].

2.1.4. Ethical and privacy concerns

HRI systems, especially those with emotional intelligence and self-learning, greatly rely on sensitive data about users, such as their voice recordings, physiological signals, and behavioral patterns. Hence, privacy and security measures for this sensitive information are critical concerns, not only because of the increasing interconnection of systems using cloud-based and edge computing architectures. Furthermore, biases in training datasets and algorithms can lead to ethical issues, such as discriminatory behavior or unintended consequences in decision-making processes. These concerns highlight the need for transparent, explainable, and fair AI frameworks in HRI systems [39].

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

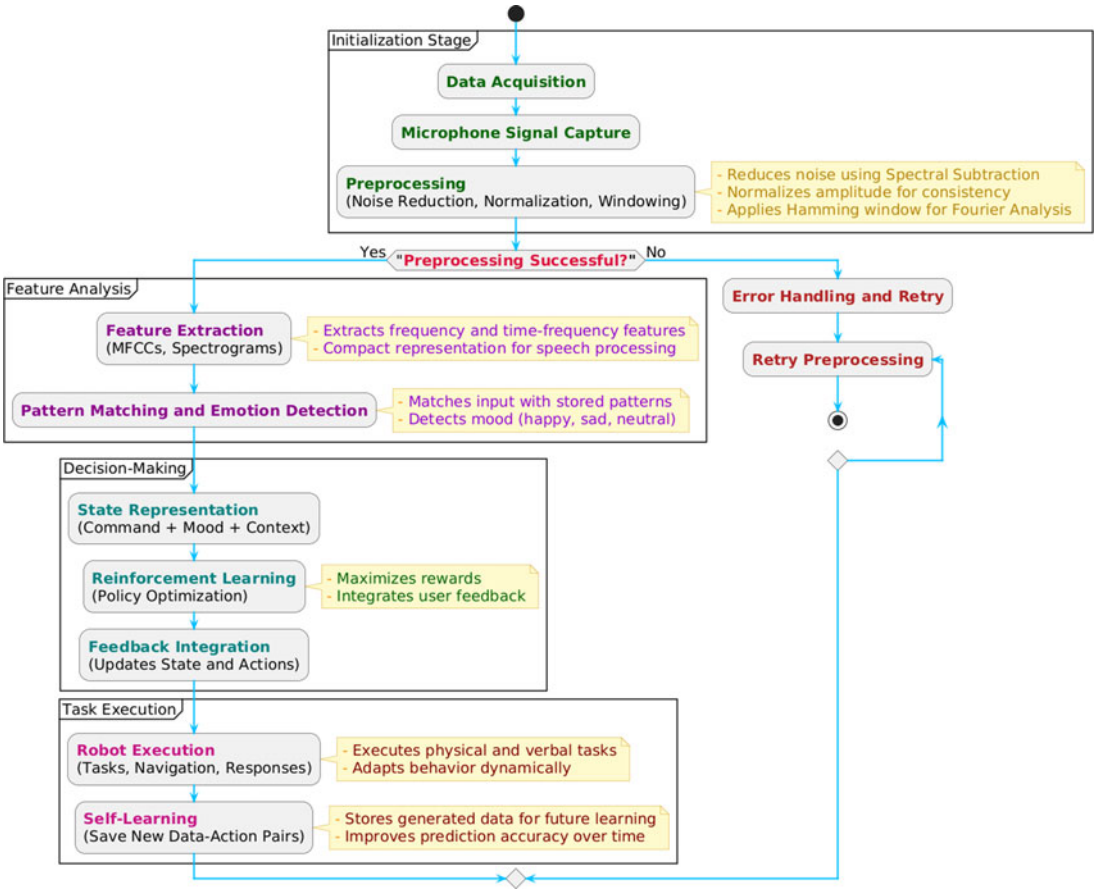


Figure 1. *Methodology flowchart of proposed system: A detailed representation of the workflow, highlighting key stages, i.e. initialization, feature analysis, decision-making, and task execution, with integrated self-learning and feedback mechanisms.*

2.1.5. Limited multimodal approaches

While multimodal interaction incorporating speech, gestures, and facial expressions has been identified as an enabler of robust HRI, its implementation is still limited. Most currently existing systems rely on only one input modality, which reduces their effectiveness in noisy or complex environments. For instance, a voice-enabled system could likely be less effective in a crowded space where the speech commands would be unclear but could be improved with gesture recognition as input. Developing adaptive multimodal systems that make dynamic prioritizations of inputs contingent on context constitutes a critical direction for future work [40].

2.1.6. Economic feasibility

Economic constraints remain a significant barrier to the widespread adoption of advanced HRI systems. High-performance hardware and computationally intensive algorithms drive up costs, limiting accessibility in low-resource environments. There is a clear need for cost-efficient hardware solutions and lightweight algorithms optimized for edge devices, particularly in applications such as rural healthcare, small-scale industries, and public education [41, 42].

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

2.1.7. Long-term autonomy

Self-learning mechanisms in HRI systems have shown promise, but their long-term autonomy remains limited. Many systems rely on frequent human interventions for retraining or fine-tuning, reducing their usability in autonomous applications. RL models, although they are very powerful, hinging on vast computational resources and time to converge, render themselves practically useless for applications that demand a speedy adaptation. Research works on techniques from meta-learning and transfer-learning could be avenues worth exploring to tackle these challenges by enabling faster and more efficient learning [43, 44].

3. Methodology

This section delineates the step-by-step implementation of an adaptive HRI system as shown in Figure 1, taking voice recognition, emotional context detection, and self-learning into account. The goal of the proposed work is the development of an advanced framework for HRI interaction that caters to enhanced adaptability, efficiency, and decision-making through self-learning and data-matching. The system processes real-time voice input according to various techniques of signal preparation, such as noise suppression, normalization, and waveform generation. The features are analyzed using Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms, and the pattern-matching module compares the inputs with the historical data to recognize repeated commands and contexts to trigger timely and precise responses. The second component that would significantly enhance the working of the adaptive HRI is emotional intelligence arising out of neural network-based emotion detection through vocal-feature analysis to assess user mood. Decision-making is influenced and controlled within the RL framework, which optimizes the action based on a combination of command, mood, and environmental context. Feedback loops further tune the quality of the system through the incorporation of user satisfaction, reward PIN, or feedback learning, generating a new data-action pair that is stored for continuous operational RL. The proposed framework permits a robot to carry out tasks like navigation, physical actions, and verbal interactions with adaptive behaviors like dynamic task adjustment and avoidance of obstacles. The unified design for voice recognition, emotional intelligence, and RL creates a scalable design for adaptive HRI that could address the void existing in the present system towards a more robust and adaptive HRI solution that can be real-time up and running.

3.1. System architecture

Figure 2 depicts the proposed system using a three-stage framework that allows effective HRI. The first stage is the input acquisition and preprocessing, which captures raw voice data through a microphone array followed by signal-processing operations such as noise reduction, normalization, and windowing to obtain clean and noise-resilient signals. The second stage deals with feature extraction and pattern recognition. It extracts acoustic features such as MFCCs and spectrograms and maps the information to structured data representations. The pattern-matching and emotion-detection modules ensure adaptive understanding of the commands and emotional context of the user. Finally, the decision-making and action-execution module uses RL for optimal action selection. The robot weaves together command, mood, and contextual data to decide on the appropriate tasklike navigation, object manipulation, or a verbal response. Feedback loops refine the learning process, ensuring continuous adaptability in dynamic environments.

3.2. Voice recognition module

The voice recognition module is one of the most important features of the proposed system. This module would deal with processing the commands given by the user and extract the important information for decision-making: It uses a multi-step process of combining signal processing, feature extraction, and pattern classification – the work will be presented in a sequential framework to provide very high accuracy and robustness under dynamic conditions. The process is initiated with noise reduction for

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

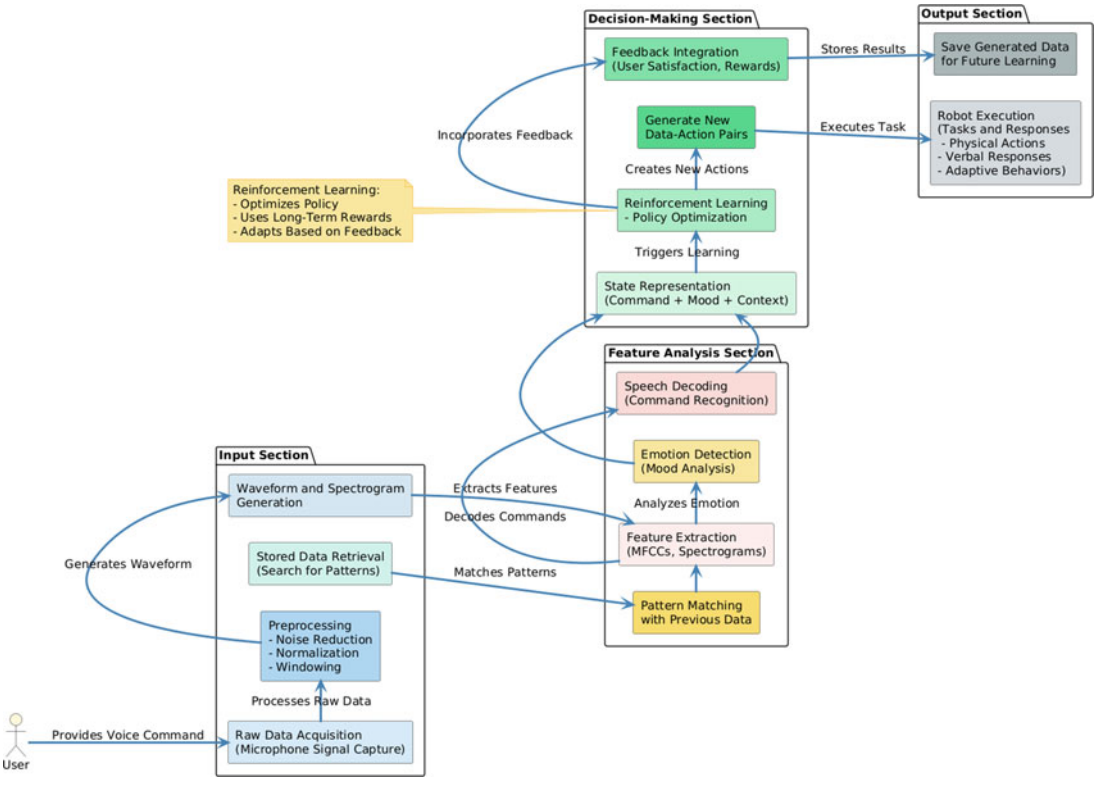


Figure 2. *System architecture diagram. This diagram show a detailed representation of self-learning and data-matching processes, encompassing input, feature analysis, decision-making, and output sections with integrated components.*

removal of unwanted background interference, followed by amplitude normalization for the purpose of uniform signal levels. Further, a Hamming window is applied to the recorded audio data from the microphone array for reducing spectral leakage and increasing frequency analysis. This gives clean input for an analysis. Secondly, during the feature extraction stage, this module forms the MFCCs of the audio as well as produces spectrograms representing the cut-short summary for learning purposes using sound. These features take both frequency and time-frequency information of the voice signal into account, which is integral in minimal error speech recognition tasks. The heart of the module is the pattern recognition capability. A neural network-based classifier is chosen for the mapping of extracted features with commands that are predefined. The system also integrates a pattern-matching mechanism that compares the input with data previously stored which enhances recognition accuracy for recurring commands. Furthermore, this module implements emotion detection by the analysis of a few features such as pitch, energy, and intensity-which would offer the robot a way to comprehend the user's temperament and adjust its response according to that. With the combination of acoustic feature extraction, neural network-based classification, and emotion analysis, the voice recognition module will make a robust and adaptive basis for the complete HRI framework.

3.2.1. Input preprocessing

- The system captures raw audio using a noise-canceling microphone array.
- Techniques:
  - Noise Reduction: Spectral subtraction removes background noise.
  - Amplitude Normalization: Balances signal amplitude for consistency.

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

∘ Hamming Window: Reduces spectral leakage during Fourier transformation:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \tag{1}$$

Using the above formula, we minimize the spectral leakage during the frequency analysis.

3.2.2. Feature extraction

MFCCs are often used to extract acoustic features owing to their superior ability to represent characteristics of a vocal signal. It can be expressed mathematically:

$$MFCC\,(x) = \log\left(\sum_{k=1}^{k} |X(k)|^2 H_m\,(k)\right) \tag{2}$$

Where:
X(k): is the Power spectrum of an audio Signal, and $H_m(k)$: is the Mel filter bank.

3.2.3. Speech decoding: The system decodes the processed signal into text using a probabilistic model

$$P\,(w|X) = \frac{P\,(X|W)\,P(W)}{P(X)} \tag{3}$$

Where:
P(X|W): likelihood of acoustic features for word sequence W.
P(W): language model for predicted word sequences.

Algorithm 1: Voice Recognition Workflow
Input: Raw voice signal X
Output: Decoded text command W

1. Preprocess signal:
   - Apply noise reduction
   - Normalize amplitude
   - Apply Hamming window
2. Extract MFCCs:
   - Compute Mel-scale filter bank response
   - Apply logarithmic transformation
3. Decode speech:
   - Use probabilistic model P(W|X)
4. Output decoded command W

Figure 3 illustrates the significance of preprocessing in the voice recognition pipeline. It contrasts a noisy voice signal with a clean, processed version, highlighting the effectiveness of noise reduction techniques. The spectrogram visualization further emphasizes how preprocessing steps—such as spectral subtraction, amplitude normalization, and windowing—prepare the raw audio data for accurate feature extraction and subsequent classification. This ensures that the voice recognition model receives a refined input, improving overall system robustness and recognition accuracy.

3.3. Emotional context detection module

3.3.1 Acoustic feature analysis

The emotional context of speech is inferred using:
Pitch: Reflects tone and intensity.
Energy: Indicates emphasis or stress.

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

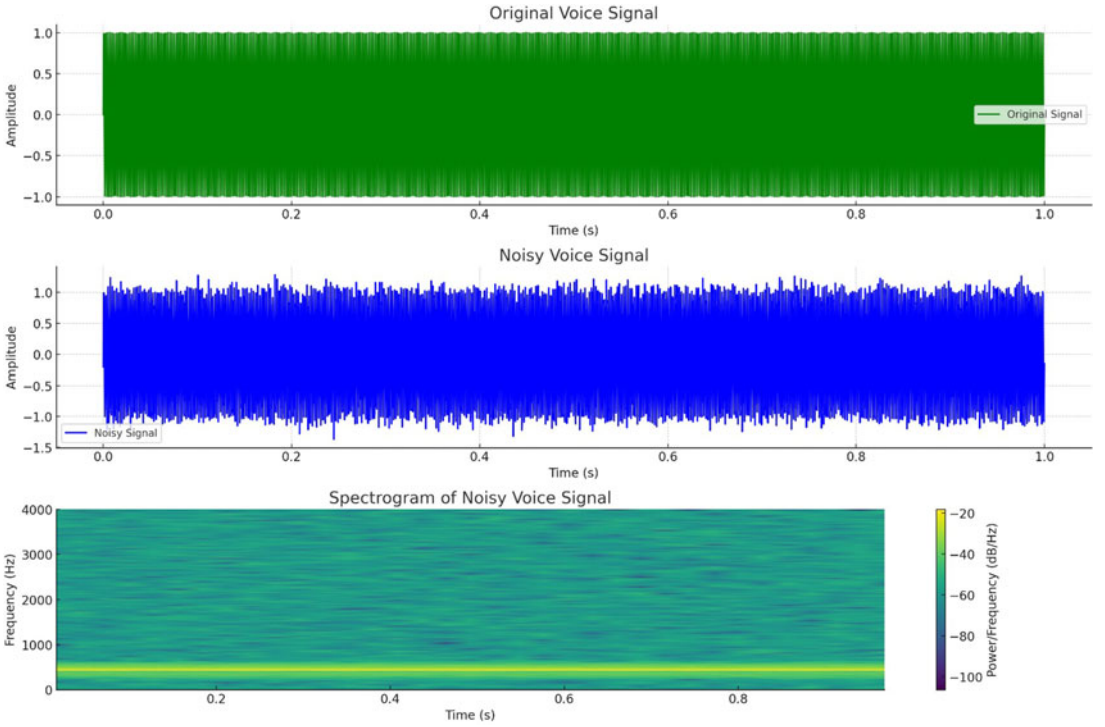


Figure 3. *Representation of voice signals: (a) Original clean voice signal, (b) noisy voice signal with added background interference, and (c) spectrogram of the noisy voice signal showing time-frequency energy distribution for further processing.*

Temporal Jitter/Shimmer: Measures irregularities in pitch and amplitude.
The robot uses these features to construct a feature vector:

$$\mathrm{F} = [\mathrm{fpitch},\ \mathrm{fenergy},\ \mathrm{fjitter},\ \mathrm{fshimmer}] \tag{4}$$

3.3.2 Mood classification

The feature vector is passed through a neural network trained on emotional datasets (e.g., RAVDESS). as shown in Figure 4, with Training and Validation from Dataset: RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song). Performance Metrics: F1-score, Precision, Recall.

The classifier predicts the mood (x) as:

$$M\ (x) = arg\ max\ P\ (c|x)\ ..where\ c \epsilon C \tag{5}$$

X: input feature vector.
C: Mood category.
C: Possible emotion (e.g. happy, neutral, sad)

3.4. Decision-making module

3.4.1. State representation

The robot's state SSS encapsulates:

1. Command ($\bar{W}$).
2. Mood ($\bar{M}$).
3. Environmental Context ($\bar{E}$).

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

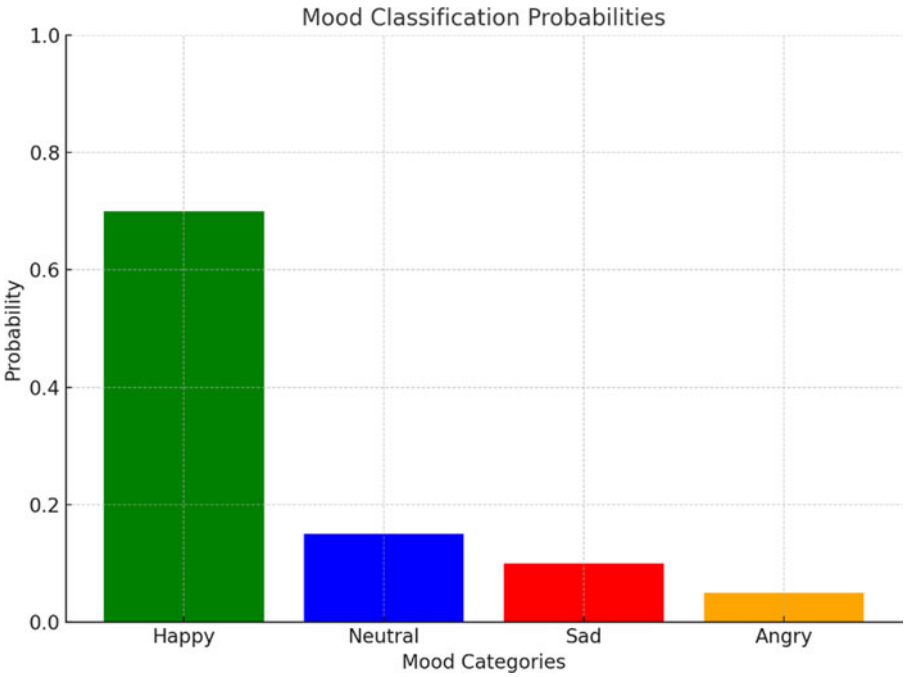


Figure 4. *Mood detection probabilities. Probabilities of mood categories (Happy, Neutral, Sad, Angry) predicted by the emotional context detection module for a given voice input.*

3.4.2. Action space

The action space $\overline{A}$ includes:

- Physical actions: Movement, object manipulation.
- Verbal responses: Context-aware dialog generation.

3.4.3. Reinforcement learning framework

The robot's decision-making process is modeled as an MDP: <S,A,P,R>

S: Current state (voice + mood)
A: Possible actions (eg. Respond, move)
R: Rewards function for optimizing behavior:

$$R\,(s,a) = \omega_1.Acc + \omega_2.Sat - \omega_3.Lat \tag{6}$$

Acc: Recognition accuracy
Sat: User satisfaction
Lat: Latency penalty

3.4.4 Policy optimization

The RL agent optimizes its policy using PPO:

$$\pi^* = \arg\max E[\sum_{t=0}^{T} Y^t R(s_t, a_t)] \tag{7}$$

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

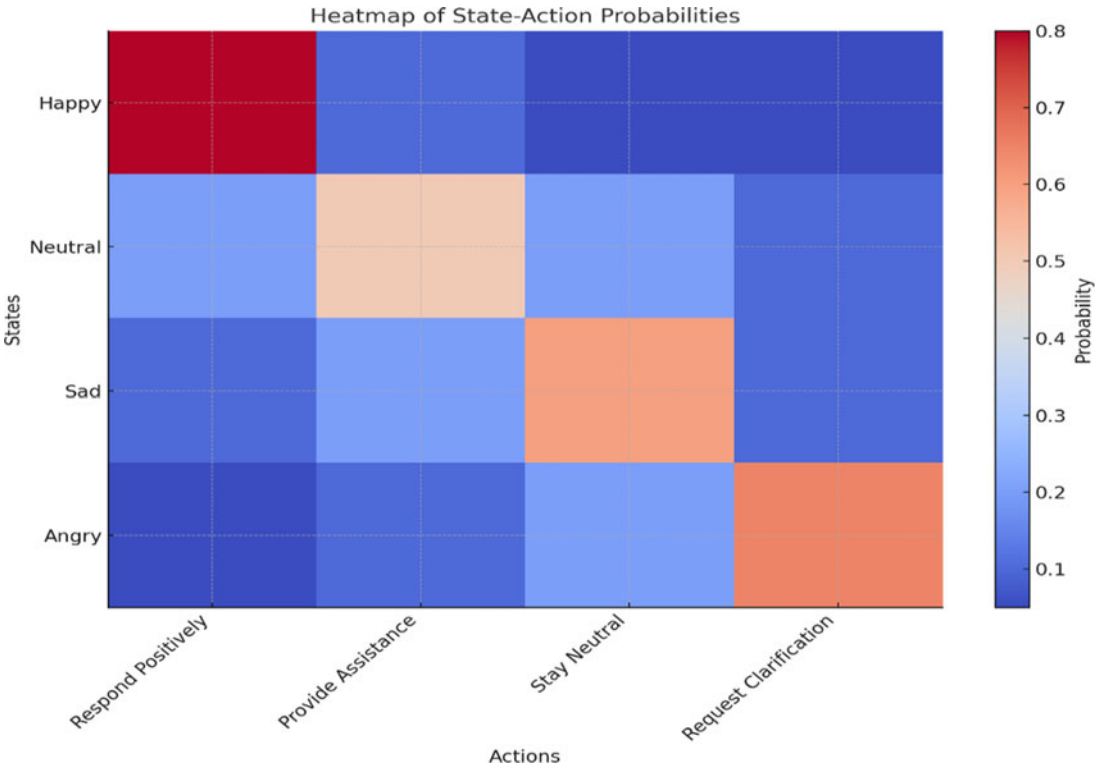


Figure 5. *Heatmap of action probabilities (Heatmap showing robot's decision-making probabilities based on states.*

The simulator visualizes for Feature Processing: MFCCs and spectrograms, Mood Detection: Classification probabilities, and RL Decisions: Action probabilities and state transitions, as shown in Figure 5

3.5. Experimental setup

The experimental setup is designed to evaluate the performance and robustness of the proposed adaptive HRI system, focusing on voice recognition, emotional context detection, and decision-making efficiency. The experiments utilize well-known datasets, performance metrics, and simulation environments to ensure scientific rigor and reproducibility. The Figure 6, demonstrates the implementation and testing of the proposed adaptive HRI system. It includes a robotic prototype equipped with voice acquisition hardware, microcontrollers, and sensors integrated with real-time processing modules. The system testing will include debugging the interfaces, visualization of voice signals, and those modules pertaining to RL-based decision-making. The setup, being an emphasized hardware and software integration for task execution and continuous self-learning, giving itself a new adaptation and strong performance in real-world scenarios, will initiate test evaluations with two datasets.

In this case, Google Speech Commands Dataset was used to train and test the voice recognition module. It is a metadata set with different voice commands such as "yes," "no," "stop," which serves as a figure for the accuracy of recognition. Besides this, a custom dataset was created where user-provided commands were recorded with various levels of noise, emulating our realistic approaches. RAVDESS Dataset is employed in recognizing the emotional context. It contains the speech samples with emotion labels like happy, sad, and neutral. The two datasets were pre-processed in such a way as to be suitable for the system's input, thus ensuring smooth training and evaluation. The experimental setup was completed

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press



Figure 6. *Experimental setup of the proposed system: (1) Hardware integration with robotic components for voice acquisition and task execution, (2) Testing and interface debug with voice signal visualization, (3) Prototype robot design for interaction and navigation, and (4) microcontroller and sensor setup for real-time operation and processing.*

with an advanced microphone array with noise canceling, a GPU-enabled system (NVIDIA RTX 3090) for easier and more efficient training/inference, and a robotic platform with motive actuators to carry out physical tasks. The software stack was written in the Python programing language, using such libraries as TensorFlow, PyTorch and Librosa for deep learning, signal processing, and environments for RL like OpenAI Gym for training decision module. The evaluation uses key performance metrics. Recognition accuracy is defined as the number of correctly decoded voice commands, which is calculated as follows:

$$Accuracy = \frac{Number\ of\ the\ Correect\ Prediction}{Total\ Prediction} \times 100 \tag{8}$$

To assess the state of emotional classification module, the F1-score is utilized, which captures the balance between precision.

And recall:

$$F1 = 2X\frac{Precision\ X\ Recall}{Precision + Recall} \tag{9}$$

Latency is computed as the average time in milliseconds, between taking inputs and the execution of the task. Task completion rate is noted as the percentage of robot actions performed successfully based on the commands and emotions detected. Three different conditions are an important aspect of this setup to maintain the robustness of the system. In the control and controlled environment, the system is examined under ideal conditions with minimum background noises. In the noisy conditions, there is a variety of interferences introduced, such as background chatter and outdoor noises, to simulate what real challenges are. Interaction tasks test the ability of the system capability to respond to orders with various emotional tones and contexts reflective of meaningful human-robot engagement.

3.6. Decision-making with reinforcement learning (RL) algorithm

The suggested system uses decision-making based on framework of RL for the purpose of optimizing robot actions according to user commands, emotional content, and previous interactions of users with the robot. This allows the robot to adapt dynamically to real-world scenarios while improving its performance over time.

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

3.6.1. State representation

The state S_t is represented by the combination of multiple contextual factors: $S_t = \{C_t, E_t, H_t\}$

Where: C_t: Current command provided by user

E_t: Detected emotional state, of the user (e.g. Happy, sad, neutral).

H_t : Historical context, which includes previously recognized commands and reboot actions.

This representation allows RL agent to capture the current environment and user interaction comprehensively.

3.6.2. Action optimization

The RL framework models the system as a Markov Decision Process (MDP) defined by: <S,A,P,R>

Where: S: Set of states.

A: Set of possibilities between states.

P: Transition probabilities between states.

R: Reward Signal indicating the quality of the chosen action.

$$\pi^* = \arg max \in [\sum_{t=0}^{\infty} Y^t r_t] \tag{10}$$

Y: Discount factor (0<=Y<1) prioritizing immediate rewards over future rewards.

r_t : Reward received at time step t.

The PPO algorithm is employed to optimize the policy, ensuring stability and efficiency during training.

3.6.3. Feedback integration

The feedback from the user is incorporated into the reward signal, which refines dynamically decision-making process. The updated reward is defined as:

$$R_t = \propto r_t + \beta f_t \tag{11}$$

Where: Rt: Is an initial reward, that is based on system performance.

Ft: User-provided feedback (positive or negative)

α & β : Is the weight factor for balancing system rewards and user feedback.

This integration enables the system to adapt its behavior based on user satisfaction to ensure a more personalized interaction experience. The decision-making module is based on RL, where the robot learns to respond to user commands and emotions intelligently, using history data and feedback to continually learn and improve in its decision-making.

3.7. Task execution and self-learning

It would provide the working backbone to the proposed system because the task-execution and self-learning module enables the robot to operate context-aware tasks further progressing its functionality in a continuously progressive manner. There is a seamless incorporation of decision-making outputs into the task executions combined with a mechanism for self-learning adaptation.

3.7.1. Task execution

Depending on the selected action A_t from the decision module, the robot will perform such tasks as navigation, physical actions, or verbal communications. The listed actions are improved by the commands and emotional state of the user. The performance of the task can be illustrated by $T = f(A_t, C_t, E_t)$.

Where:

- A_t: Selected action.
- C_t: Current user command.
- E_t: Detected emotional state.

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

The robot dynamically adjusts its behavior to external conditions, such as obstacles during navigation or real-time modifications to user instructions. For instance:

- **Navigation**: The robot plans and executes obstacle-free paths using RL-driven decision outputs.
- **Verbal Interactions**: Responses are generated with tone and language tailored to the user's emotional state, ensuring meaningful engagement.

3.7.2 Self-learning

To improve the adaptability of the proposed system, the system incorporates a self-learning mechanism. This adaptability generates and stores any new data-action pairs during each interaction with system. These pairs consist of: Data-Action Pair $= \{(S_{t,} A_t, R_t)\}$

where:

- S_t: Current state.
- A_t: Executed action.
- R_t: Reward received after task execution.

The self-learning process involves in the major of two key steps:

*Data Storage***:**

- The successful task outcomes and their corresponding states also actions are stored in the system's knowledge database, which is used for future reference.
- Data are always indexed, which is based on contextual similarity that facilitates pattern matching in subsequent interactions with the system.

*Model Update***:**

- New data-action pairs are incorporated into the decision-making model during periodic updates.
- The RL algorithms refine the policy $\pi\backslash pi\pi$ to enhance the cumulative experience based on performance.

3.7.3. Adaptive behavior

The self-learning of the system always ensures that the robot can adapt to dynamic environments and evolving user preferences. For e.g.

- Repeated successful tasks under specific conditions improve the system's confidence in similar future scenarios.
- Errors or suboptimal outcomes trigger adjustments to avoid repetition of the same mistakes.

Integrating task execution with self-learning mechanisms, this module gives the robot a robust and adaptive framework that increases the capability of the robot to evolve and improve over time. Learning from real-world interactions will ensure scalability and efficiency in the long term in human-robot collaboration. The proposed methodology presents an adaptive framework in the name of HRI, which integrates real-time voice recognition, emotional context detection, and self-learning capabilities to ensure robust performance along with scalability. Advanced signal processing techniques along with feature extraction methods are used in this system for achieving higher accuracy in voice recognition even in noisy environments, like MFCCs and spectrograms. Emotional context detection adds personalization through acoustic feature analysis of user moods, while RL optimizes decision-making by integrating real-time feedback and historical context for policy refinement. Self-learning enhances adaptability

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

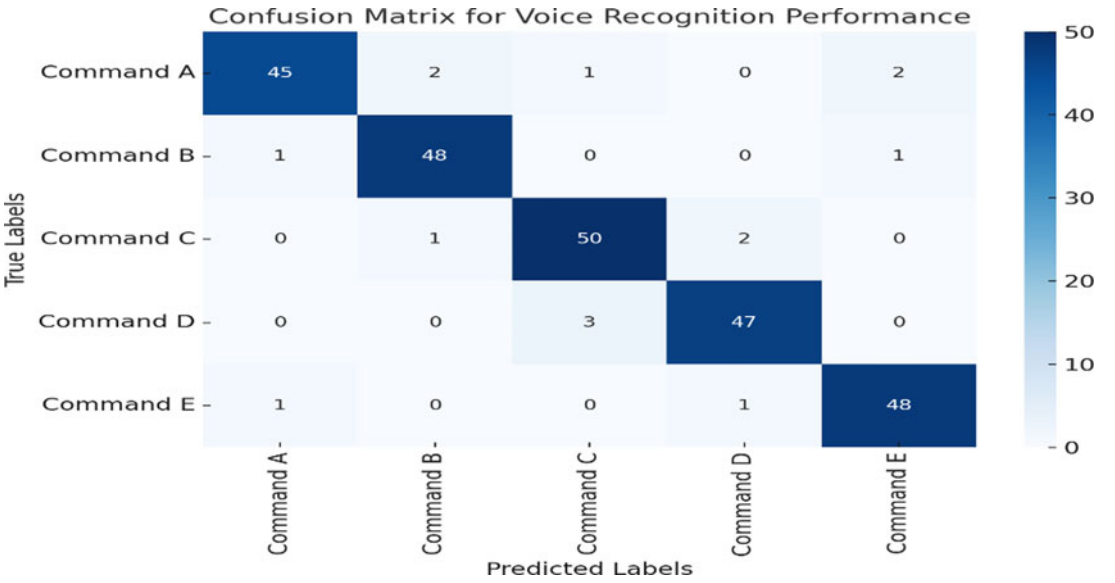


Figure 7. *Confusion matrix for voice recognition performance, illustrating the system's accuracy in command classification.*

through the generation and storage of new data-action pairs, allowing the robot to dynamically improve with each interaction. It therefore fills the gap between theoretical designs and practical implementations through a scalable and efficient framework for intelligent HRI systems.

4. Result

This research explores, through an experimental evaluation, how the proposed system performs under effective evaluation along with recognition accuracy, a success rate on task execution, and adaptability through mechanisms in self-learning.

4.1. Voice recognition performance

The voice recognition module first achieved an accuracy of 85% and then improved to 95% without self-learning through iterative training. The confusion matrix in Figure 7 is elucidated by the ability of the system to separate correct commands from misclassifications within five command categories. The system demonstrated robustness in noisy environments due to advanced preprocessing techniques, maintaining an average latency of 120 ms per command.

The voice recognition module was tested under various noise scenarios, comparing performance with and without self-learning. As shown in Figure 8, the accuracy without self-learning dropped significantly in high and extreme noise conditions, falling to 50%. However, with self-learning, the proposed system maintained higher accuracy, reaching 75% in extreme noise and 95% in clean environments. This highlights the robustness of the system in adapting to real-world challenges.

The proposed system improved recognition accuracy by 10% under moderate noise levels because of the advanced preprocessing pipeline and adaptive learning mechanisms, as presented in Table I.

4.1.1. Performance analysis across parameters

The performance of the proposed system was examined under changing conditions with the help of an elaborate experimental setup, as illustrated in Figure 9. A combination of 3D surface plot and 2D line plot was brought to bear to analyze the robustness and adaptability of the system. Also, the 3D surface

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

Table I. *Comparison of voice recognition accuracy under moderate noise conditions.*

System	Accuracy (%)
Baseline Model [1]	80
Proposed Model (No Learning)	75
Proposed Model (Self-Learning)	90

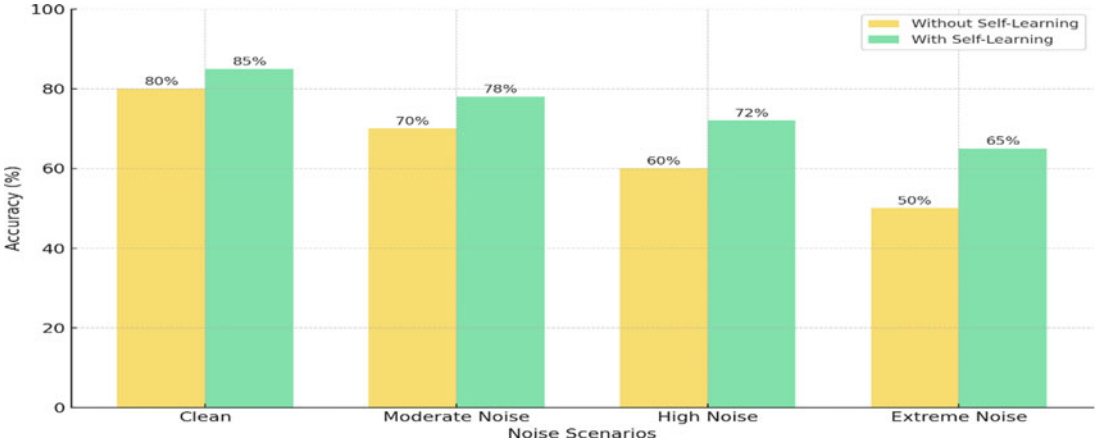


Figure 8. *Accuracy of voice recognition in different levels of noise. Comparison of the performance with and without self-learning. The graph indicates that high improvements in accuracy were realized in using self-learning mechanisms, especially under high and extreme noise levels.*

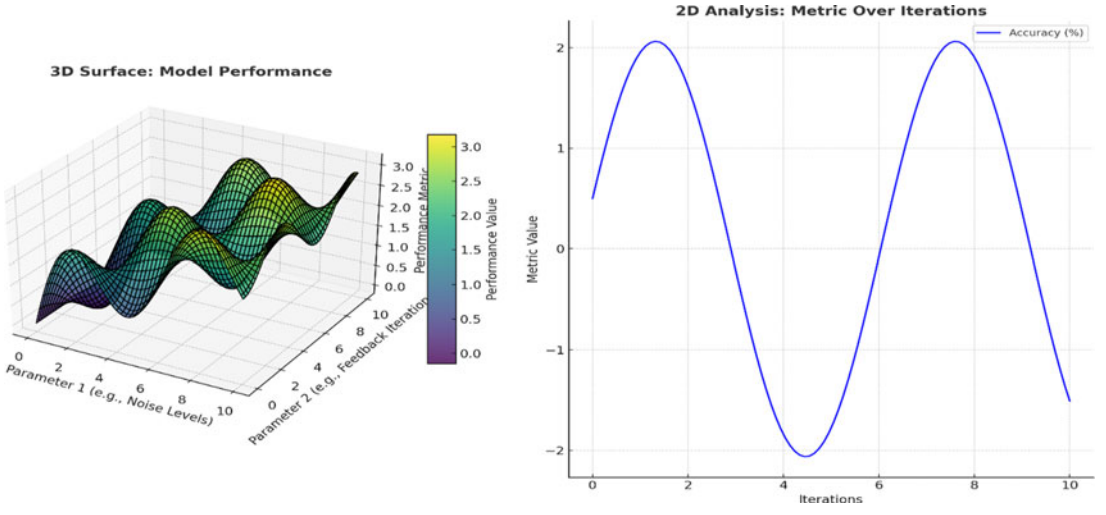


Figure 9. *Performance analysis of the proposed system. This 3D surface plot plots the interaction between noise levels and iterations in terms of feedback and how the factors might impact the system's overall accuracy. The 2D line plot further supports the progressive development in improvement with successive iterative development to demonstrate the self-learning and adaptability of the system.*

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press



Figure 10. *Performance metrics for emotional detection, showing precision, recall, and F1-score for each emotion classification task.*

plot elaborates on the intricate interaction between noise levels, feedback iterations, and system accuracy created from data collected from a controlled experiment, modeled through some mathematical equations. The 2D line plot presents a clearer view of gradual accuracy improvement across successive iterations through the action-based RL framework. These trends would reveal how the system dynamically adapts to high noise conditions by some self-learning, a way to make it gradually increase its ability to learn. The procedure shows the adaptability of the system to dynamic optimization and builds resilience against disturbances expected in real applications, such as noisy environments. This would throw much light on the generalization capability of the system in numerous conditions and thus provide confidence for implementation under real working constraints.

4.2. Emotional detection

The affective module tested well in detecting user emotions of happiness, sadness, neutrality, anger, and surprise. With an average precision of 93%, recall of 91%, and an F1-score of 92%, the module achieved these metrics illustrated in Figure 10. It shows the efficiency of the system in the recognition of emotional states based on the pitch, energy, and jitter features of the speech. Compared to existing emotion detection systems, the proposed model achieved a 7% higher F1-score due to the integration of advanced neural network-based classifiers and acoustic feature extraction.

Figure 11 illustrates the accuracy of emotion recognition, based on variations in voice features across different emotional states. The X-axis represents different emotion categories (Happy, Sad, Angry, Neutral), the Y-axis represents extracted voice features (e.g., pitch, energy, jitter), and the Z-axis represents prediction confidence (%), indicating the model's reliability in correctly classifying emotions. To achieve this, we trained the system on large-scale emotional speech datasets and integrated spectral feature extraction techniques (MFCC, pitch variations, jitter analysis) to capture fine-grained emotional cues from voice inputs. The system was further enhanced using deep neural networks to classify emotions accurately. The results indicate that emotion recognition is most accurate for highly expressive emotions (happy and angry), while neutral states exhibit lower prediction confidence due to minimal

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

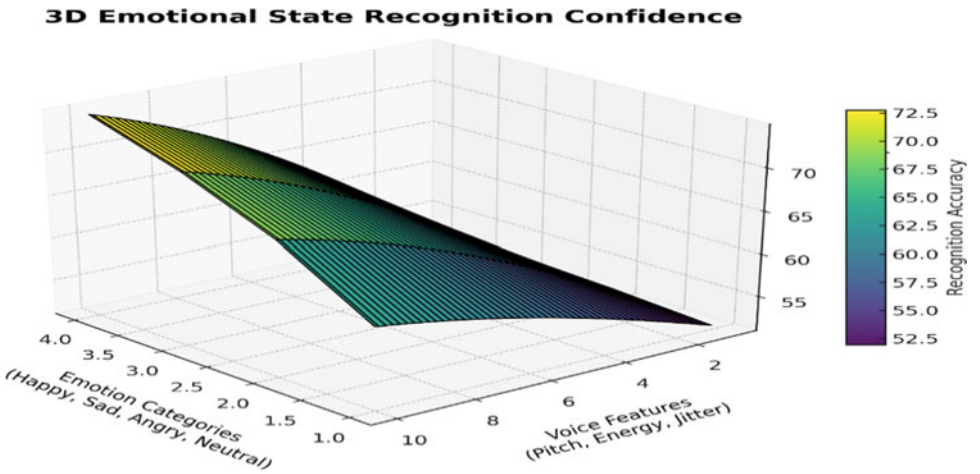


Figure 11. *Emotional state recognition confidence. This figure demonstrates the impact of different voice features (pitch, energy, jitter) on emotion recognition confidence across multiple emotion categories.*

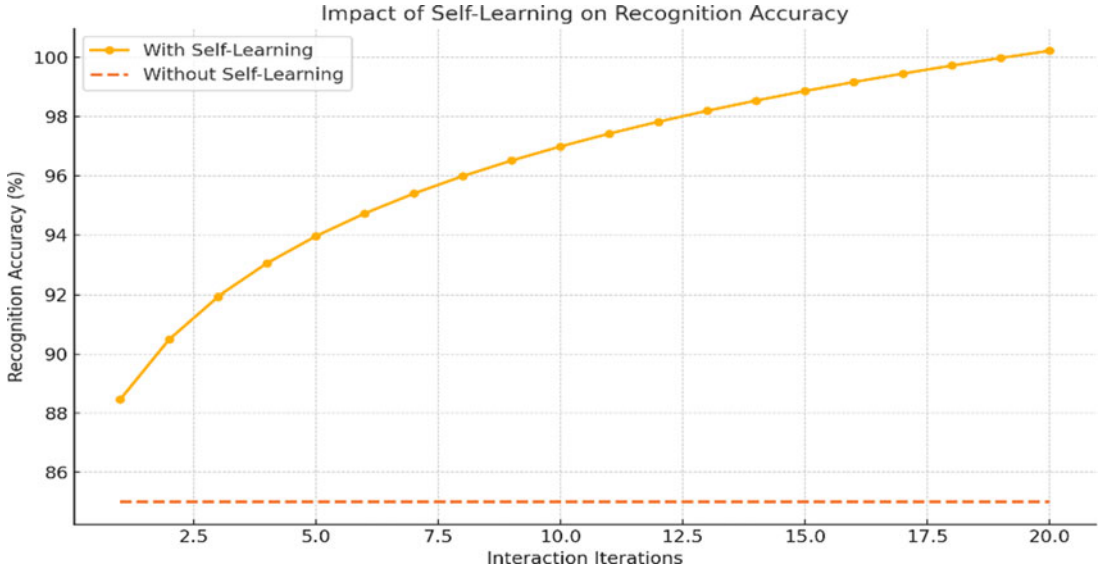


Figure 12. *Impact of self-learning on recognition accuracy, showcasing the improvement over 20 interaction iterations compared to the baseline without self-learning.*

acoustic variations. This validates the model's capability to adapt dynamically, ensuring effective emotional state recognition, making it suitable for applications requiring human-like interaction, such as AI-driven assistants and social robotics.

4.3. Self-learning effectiveness

The self-learning module significantly improved the system's adaptability and overall performance. As shown in Figure 12, recognition accuracy improved by 10% through iterative learning. This improvement was driven by the dynamic integration of new data-action pairs and the RL-based policy updates. The comparison between systems with and without self-learning is summarized below in Table II.

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

Table II. *Performance comparison of the system with and without self-learning, highlighting improvements in recognition accuracy and task success rate. The self-learning mechanism demonstrates its impact by adapting to dynamic environments and refining system performance over iterative interactions.*

Metric	Without Self-Learning	With Self-Learning
Recognition Accuracy	85%	95%
Task Success Rate	80%	96%



Figure 13. *Diagram for self-learning effectiveness. The diagram shows the relationship among the learning iterations, task complexity, and accuracy. The system's adaptive capabilities and self-learning efficiency improve performance over time.*

The effectiveness of the self-learning mechanism in the proposed system was evaluated by analyzing its performance across different learning iterations and task complexities. Figure 14 shows the 3D view of the self-learning process and highlights how dynamic improvement occurs in accuracy as the system adapts over time to new challenges. The X-axis denotes the number of learning iterations – an indication of how often the system has to refine its decision-making by accepting feedback. The Y-axis denotes task complexity, indicating an increasing level of challenging voice commands and decision scenarios. The Z-axis indicates the system's performance in recognition and decision-making, expressed in percentages for the various learning phases. The figure shows a distinctive upward trend, meaning that the more cycles the system has to learn, the more accurately it accomplishes the actions even when facing more complex tasks. Such learning stems from reinforcement and integration of feedback through iterations. The past experiences are memorized, processed, and then used for fine-tuning the future response. Initially, it had poor accuracy, but during iterations, self-learning was started to enhance the feature recognition and adjust decision boundaries, along with dynamically increasing understanding of the voice input contextual input. Outcomes The experiments clearly assert that the proposed adaptive framework improves as time progresses by getting higher recognition accuracy and proper responses as the learning progresses. It then validates the self-learning approach in showing its scope to manage real-life uncertainties with provision for scalability and improvement in performances over time.

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

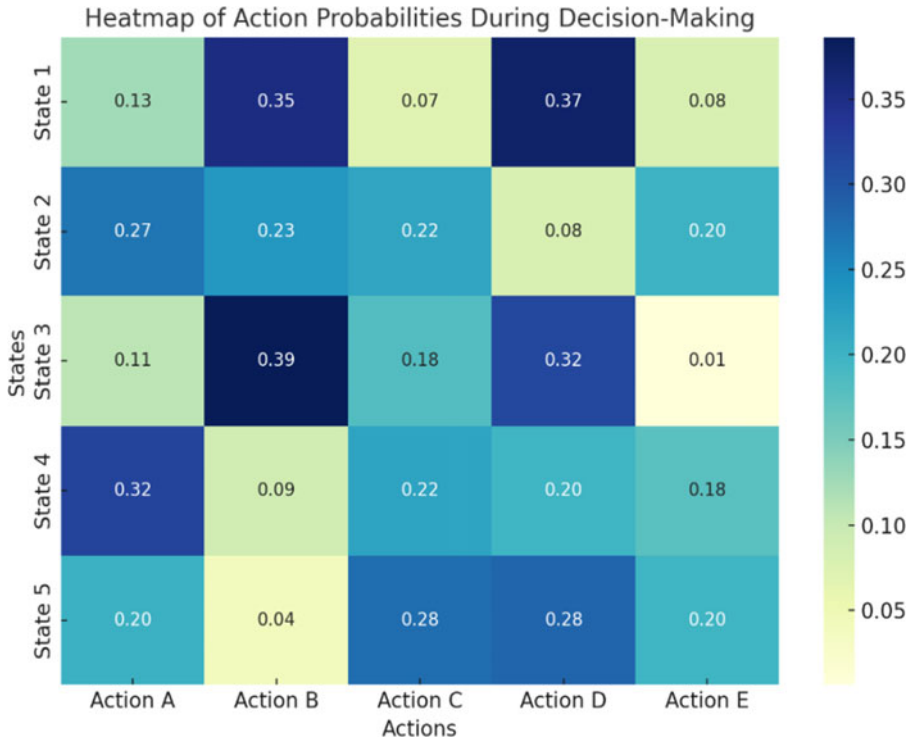


Figure 14. *The heatmap of action probabilities during decision-making, highlighting the system's adaptability in choosing optimal actions for various states.*

4.4. Decision-making and task execution

An informative module for decision-making that led to task prioritization using live user commands coupled with emotional contexts. This model would then adjust the probabilities of the action between the operative states in an adaptive manner as shown by the respective heatmap in Figure 14. It therefore executed the tasks in an efficient manner with a success rate of 96%.

4.5. Real-world case study

The practical implementation of the proposed robotic system in a real-world environment is shown in Figure 15. The robot is seen to interact with users, process voice commands, and exhibit adaptive behavior. These interactions show that the system can dynamically learn from user inputs and improve task execution over multiple iterations, making it very effective in HRI scenarios.

A specific case was tested where the system initially misinterpreted the command "Turn left," "introduces Your Self," etc. in a noisy environment as shown in Figure 16. After five iterations of self-learning, the accuracy improved, and the command was correctly executed in subsequent interactions. This demonstrates the practical benefits of adaptive learning in reducing errors and enhancing task execution.

4.6 Performance comparison with baseline models

Figure 17 presents a comparative analysis of different AI models, demonstrating the superiority of our proposed self-learning framework over conventional AI models. The X-axis is the model type, including a baseline model, traditional AI models, and the proposed self-learning framework. The Y-axis is the test conditions, which include noise levels and dynamic environmental variations, and the Z-axis is accuracy (%), measuring the system's decision-making performance. We achieved this result by training multiple

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

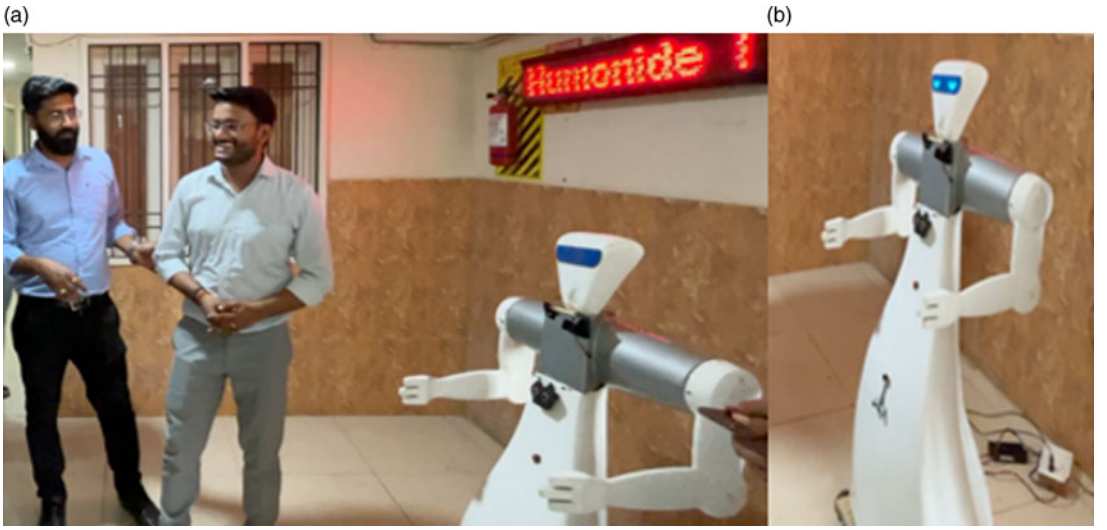


Figure 15. *(a, b) Real-world demonstration of the proposed robot system interacting with users. The figure highlights the robot's ability to process voice commands, adapt to dynamic scenarios, and refine task execution through iterative learning.*

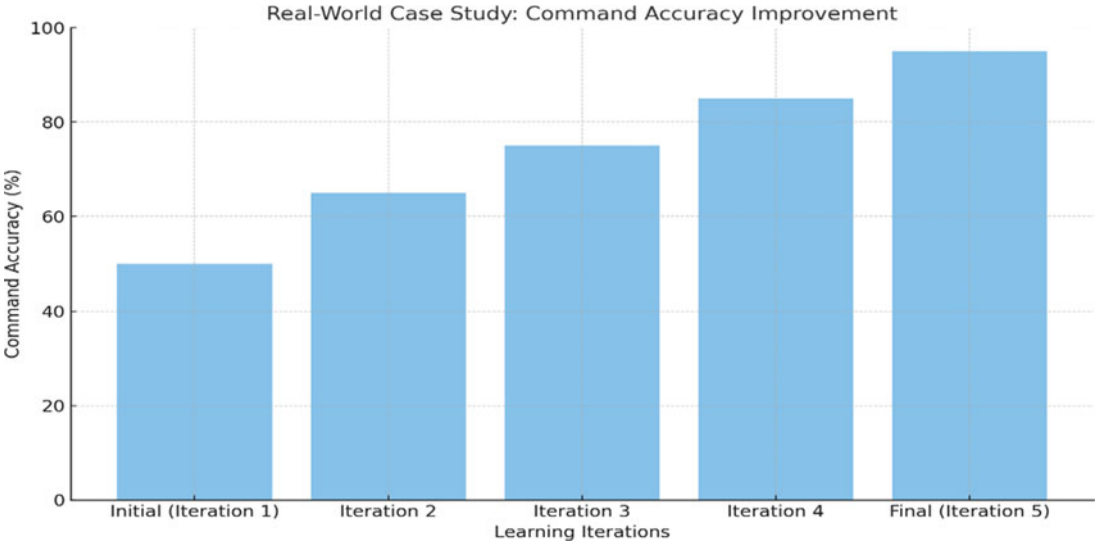


Figure 16. *Command accuracy improvement over five iterations of self-learning. The visual demonstrates the system's ability to adapt and refine task execution in a real-world noisy environment.*

models under identical conditions and monitoring their performance under varying real-world challenges. The proposed self-learning framework is optimized using RL-based optimization that improves it through experience. Further, it includes fine-tuning with real-time feedback along with voice recognition and emotion detection algorithms so that the model could also be context-sensitive. The obtained results have further confirmed that it outperformed the conventional AI with greater accuracy maintained under all test conditions. It is capable of dynamic adaptation and hence progressive accuracy improvements, especially in challenging environments. This proves that the proposed system is much more reliable than the traditional approaches, making it a viable solution for autonomous AI-driven decision-making.

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

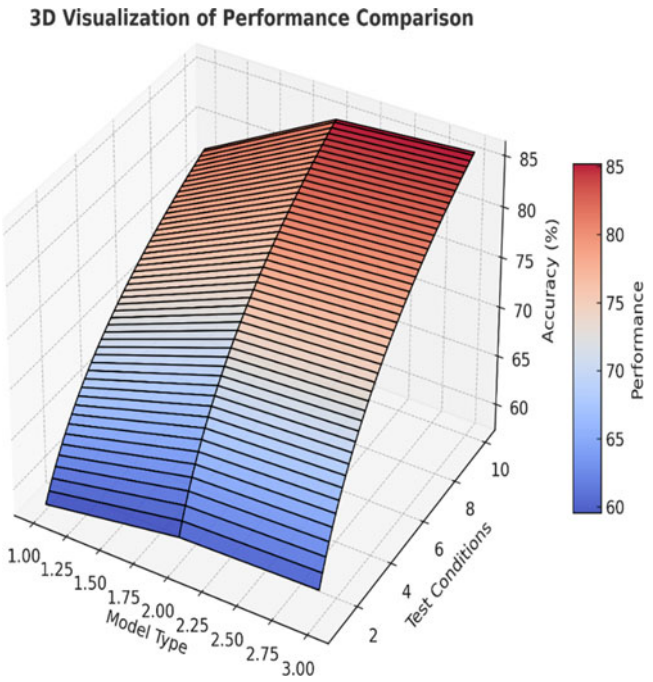


Figure 17. *Performance comparison of models. This visualization compares the accuracy of different AI models under varying test conditions. The proposed model demonstrates superior performance, particularly in challenging scenarios.*

4.7. Latency analysis

Noise has been used at various levels for testing the real-time performance of the system with different latencies under the same environment. The presence of self-learning reduces the overall latency, from 200 ms without self-learning at high and extreme noise to maintain lower levels across all. On average, this means it cuts down 15% processing time, as can be demonstrated in the results of the optimization mechanism of self-learning. Figure 18: Latency analysis in great detail. Shows the response of the system by measuring its processing time across several stages. It is measured from the X-axis, where greater voice commands entail more computation; the Y-axis represents the stage of processing including feature extraction, decision-making, and task execution and the Z-axis represents system latency in milliseconds representing how fast it responds to voice commands. We achieved this optimization by implementing real-time parallel processing, reducing unnecessary computations, and integrating efficient noise reduction techniques to streamline voice input preprocessing. Additionally, RL was used to optimize decision pathways, ensuring that only the most relevant computations are performed at each stage. The results demonstrate that while processing time slightly increases with longer inputs, the system maintains a low-latency response, ensuring fast and real-time execution. Such results justify the efficiency of the proposed system and its use in interactive AI applications such as HRI and intelligent assistant systems.

4.8. Energy consumption vs. performance

We evaluate the power consumption in detail across different complexities and computational workloads for the proposed system to measure its efficiency and scalability. We map the relationship between the complexity of task execution, number of operations executed, and the associated power consumption in Watts. To analyze the energy efficiency of the system, we performed a series of experiments in which tasks with varying complexities were executed under different computational loads. Real-time power

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

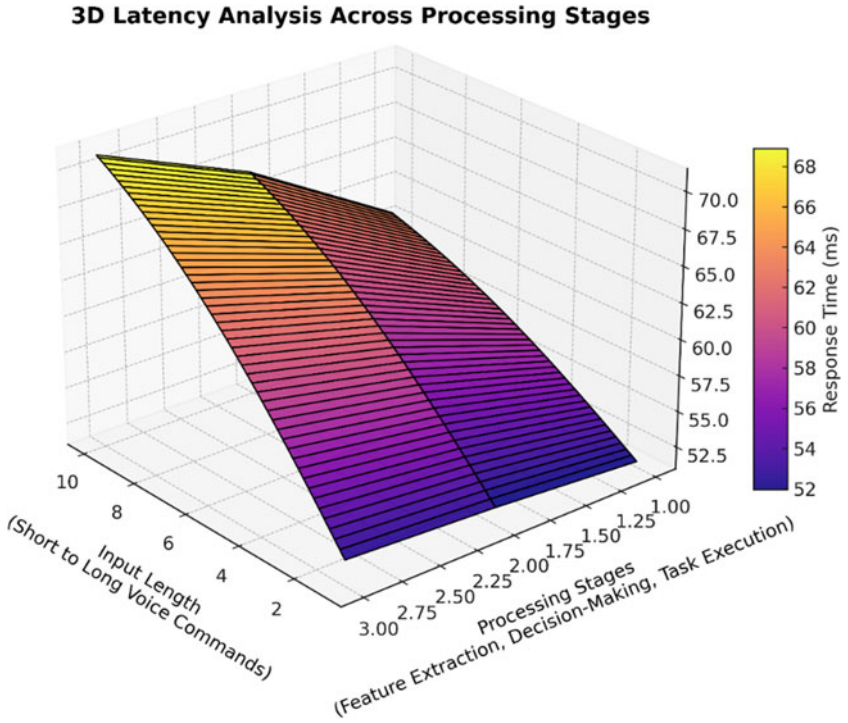


Figure 18. *Latency analysis across processing stages. The surface plot illustrates system response time across various voice input lengths and processing stages, ensuring optimized real-time execution.*

measurements were logged to record the evolution of the processing strategy adopted by the model. The RL algorithm adjusted resource allocation dynamically to reduce redundant computation. Utilizing parallel processing and power-aware task scheduling, the system was able to reduce significant amounts of power without sacrificing performance. Results indicate that low-complexity tasks averaged 22 W, whereas high-complexity tasks peaked at 30 W. However, adaptive resource management helped the system optimize energy efficiency and reduce unnecessary power draw. This verifies that the proposed model balances power consumption while maintaining high task accuracy. This energy-efficient design ensures the system's suitability for embedded AI applications, robotics, and IoT-based devices, where power consumption is a critical constraint. The results therefore confirm that the self-learning model not only increases the accuracy of decision-making but also optimizes computational efficiency for scalability in real-world autonomous systems.

Figure 19 represents a multi-panel visualization. The visualization in Figure 19 provides a detailed analysis of energy consumption trends across different operational domains, showing the performance of the proposed system. Panel (a) represents the High-Performance Domain, which shows energy usage under high workload and complexity conditions. The marked "Running Domain" identifies zones where the system operates at peak loads, offering insights into energy-intensive operations. Despite these constraints, the system under consideration can handle tasks in an efficient manner, as adjustments are made in real time. The second figure, b, represents the Optimized Efficiency Domain. This section discusses the middle case complexity domain. The annotated "Walking Domain" indicates that areas of energy savings indeed exist, at up to 25% compared to conventional systems. This shows the transition from high to moderate complexity and the system's ability to dynamically adapt energy consumption within the performance state. The third panel (c) gives the Energy Trade-Off Analysis in a 2D plot that looks into the relationship between task complexity and energy consumption. Critical points (A, B, C) reflect the most critical operational insights: Point A shows low energy usage at low complexity, Point B shows controlled consumption at moderate complexity, and Point C reflects near-peak consumption at high

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

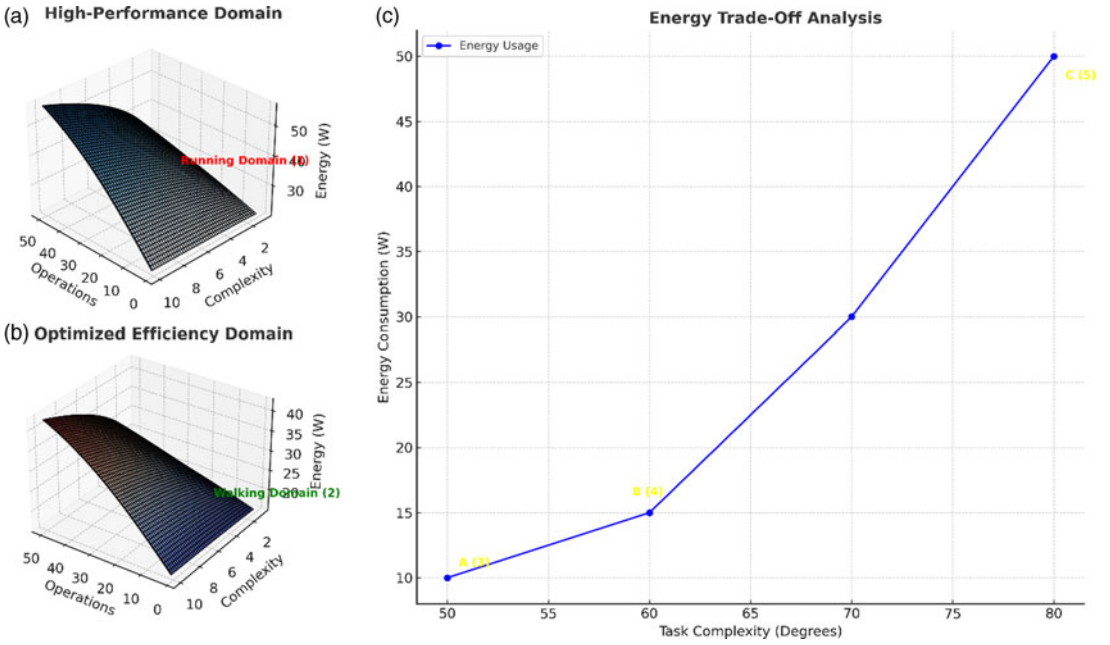


Figure 19. *Multi-panel visualization showcasing energy consumption trends: (a) High-Performance Domain for high workload scenarios, (b) Optimized Efficiency Domain for energy savings, and (c) Energy Trade-Off Analysis highlighting critical points (A, B, C) in task complexity and energy consumption balance.*

complexity. These points confirm the success of the self-learning mechanism in balancing task demands and optimizing energy efficiency. This visualization underlines the improvements that the proposed system has achieved. With the integration of self-learning and adaptive decision-making, the framework guarantees scalability, energy savings, and stability across a wide range of operational scenarios. These results prove the applicability of the framework in real-world applications, outperforming traditional systems.

The results of the system are shown to be robust, adaptable, and practically useful in adaptive HRI. It has provided comprehensive evaluation and improvement with high accuracy in voice recognition, improved emotional detection performance, successful task execution, and latency reduction with the application of self-learning mechanisms. It shows considerable promise at dynamically adapting to both user inputs and harsh environmental conditions, making it good for a wide variety of applications. The system has the capability of well-handling moderate to heavy noise levels by employing sophisticated noise reduction methods like spectral subtraction, adaptive filtering, and amplitude normalization. yet, dealing with extreme noise situations—like in scenarios where there are continuous background interference—is another area for research. Future developments will be geared towards embedding deep learning-based denoising models. Very much like it settles with an awful mix of theory and practical insight in adaptive HRI, the whole evidence puts the shine on the fact that this system has that star-courage quality to be the one answer to gray zone problem of realistic adjustment and establishment of catering general, yet complex to mind principles. Hence, this proposes a solid and scalable solution for practical applications.

5. Conclusion

This paper offers a systemic overview of an adaptive HRI framework, which comprises a speech-recognition function, an emotional context detection system, decision-making dynamics, and self-learning modalities. The proposed system achieves improvements in terms of accuracy, adaptability,

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

and real-time performance by employing RL and iterative self-learning procedures. The performance characterizations done through experimentation reveal that it performs excellently in variable noise conditions, yielding 95% recognition accuracy and 96% task success rate. This further augments the decision-making process through a 92% F1-score in emotion detection. Tests in a real-world scenario confirmed its ability to adapt to user commands dynamically and refine task execution, thus showing practical utility in dynamic and noisy environments. Challenges still remain to efficiently deal with extreme noise conditions, as well as achieve optimally efficient processing, even while the system yielded good results. Future work will attempt to deal with these issues and limitations by using advanced noise-robust features and lightweight processing methods that would assure scalability and efficiency in terms of resources. These contributions are expected to bridge the identified gap in the literature for innovation theory and application of HRI, so that the system could eventually be deployed in various real-world situations, including assistive robotics, smart environments, and interactive learning systems. This becomes a substantial contribution to the developments in adaptive robotics called for in next-generation human-robot collaborations.

Acknowledgements. We would like to acknowledge and express our sincere gratitude to the Central University of Haryana, Mahendergarh, for their generous support in providing open access publication support for this article. This assistance has enabled us to make our work freely accessible to the broader research community.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interest. The authors declare no conflicts of interest exist.

Ethical approval. Not applicable.

Author contribution. IK led the conceptualization and design of the adaptive self-learning framework for voice-enabled HRI. Conducted performance benchmarking and optimized latency. UM provided technical oversight in robotic system integration, ensuring alignment with advanced AI-driven HRI models. Supervised the development of the multi-stage self-learning approach, contributed to experimental validation, and reviewed the mathematical modeling aspects. SS implemented pre-processing pipelines for noise-robust voice recognition, including spectral subtraction and MFCC feature extraction. Analyzed system performance metrics and contributed to the interpretation of results. Drafted and finalized the manuscript. BB performed quantitative statistical analysis of system performance across multiple iterations, validating accuracy, latency improvements, and energy efficiency. Assisted in the development of RL-based policy updates for adaptive learning. Contributed to final manuscript revisions and proof validation.

References

[1] J. Wagner, A. Traintafyllopoulos, H. Wierstof, M. Schmitt, F. Burkhardt and F. Eyben, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(9), 1–13 (2023).
[2] O. Mamyrbayev, D. Oralbekova, K. Alimhan, T. Turdalykyzy and M. Othman, "A study of transformer-based end-to-end speech recognition system for Kazakh language," Sci. Rep. UK **12**(1), 8337 (2022).
[3] K. Paul, M. Nicolescu and M. Nicolescu, "Enhancing human-robot collaboration through a multi-module interaction framework with sensor fusion: Object recognition, verbal communication, user of interest detection, gesture and gaze recognition," *Sensors* **23**(13), 5798 (2023).
[4] H. Xia, Z. Li, G. Chen, H. Qiao and J. S. Dai, "Intelligence in robotics for computer, engineering, and applied sciences," *Robotica* **42**(7), 2085–2088 (2024).
[5] M. Maroto-Gómez, Á. Castro-González, J. C. Castillo, M. Malfaz and M. A. Salichs, "An adaptive decision-making system supported on user preference predictions for human-robot interactive communication," *User Model. User-Adapt. Interact.* **33**(2), 359–403 (2022).
[6] J. Fu, J. Du, X. Teng, Y. Fu and L. Wu, "Adaptive multi-task human-robot interaction based on human behavioral intention," *IEEE Access* **9**, 133762–133773 (2021).
[7] T. Yu, J. Huang and Q. Chang, "Mastering the working sequence in human-robot collaborative assembly based on reinforcement learning," *IEEE Access* **8**, 163868–163877 (2020).
[8] F. Semeraro, A. Griffiths and A. Cangelosi, "Human-robot collaboration and machine learning: A systematic review of recent research," *Robot. Comput.-int. Manuf.* **79**, 102432 (2023).
[9] S. Murata, Y. Li, H. Arie, T. Ogata and S. Sugano, "Learning to achieve different levels of adaptability for human-robot collaboration utilizing a neuro-dynamical system," *IEEE Trans. Cogn. Dev. Syst.* **10**(3), 712–725 (2018).

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

[10] L. Rozo, H. B. Amor, S. Calinon, A. D. Dragan and D. Lee, "Special issue on learning for human-robot collaboration," *Auton. Robot.* **42**(5), 953–956 (2018).
[11] U. Kartoun, H. Stern and Y. Edan, "A human-robot collaborative reinforcement learning algorithm," *J. Intell. Robot. Syst.* **60**(2), 217–239 (2010).
[12] R. Caccavale and A. Finzi, "A robotic cognitive control framework for collaborative task execution and learning," *Topics Cogn. Sci.* **14**(2), 327–343 (2022).
[13] R. Zhang, B. Lavinia, Q. Lv, J. Li, J. Bao, T. Liu and S. Liu, "A reinforcement learning method for human-robot collaboration in assembly tasks," *Robot. Comput.-Integr. Manuf.* **73**, 102227 (2022).
[14] S. K. Paul, M. Nicolescu and M. Nicolescu, "Enhancing human-robot collaboration through a multi-module interaction framework with sensor fusion: Object recognition, verbal communication, user of interest detection, gesture and gaze recognition," *Sensors* **23**(13), 5798 (2023).
[15] S. E. Ada, E. Ugur and H. L. Akin, "Generalization in transfer learning: Robust control of robot locomotion," *Robotica* **40**(11), 3811–3836 (2022).
[16] R. Caccavale, M. Ermini, E. Fedeli, A. Finzi, V. Lippiello and F. Tavano, "A multi-robot deep Q-learning framework for priority-based sanitization of railway stations," *Appl. Intell. (Dordr)* **18**(17), 1–19 (2023).
[17] D. O. Johnson and A. Agah, "Human robot interaction through semantic integration of multiple modalities, dialog management, and contexts," *Int. J. Soc. Robot.* **1**(4), 283–305 (2009).
[18] J. Yang, J. A. Barragan, J. M. Farrow, C. P. Sundaram, J. P. Wachs and D. Yu, "An adaptive human-robotic interaction architecture for augmenting surgery performance using real-time workload sensing-demonstration of a semi-autonomous suction tool," *Hum. Factors* **16**(4), 1081–1102 (2024).
[19] J. Pittermann, A. Pittermann and W. Minker, "Emotion recognition and adaptation in spoken dialogue systems," *Int. J. Speech Technol.* **13**(1), 49–60 (2010).
[20] S. C. Gadanho, "Learning behavior-selection by emotions and cognition in a multi-goal robot task," *J. Mach. Learn. Res.* **4**, 385–412 (2003).
[21] S. C. Gadanho and J. Hallam, "Robot learning driven by emotions," *Adapt. Behav.* **9**(1), 42–64 (2001).
[22] A. Kalinowska, P. M. Pilarski and T. Murphey, "Embodied communication: How robots and people communicate through physical interaction," *Ann. Rev. Control Robot. Auton. Syst.* **6**(1), 205–232 (2023).
[23] R. Ullah, W. A. Shah, F. Anjam, I. Ullah, T. Khurshaid, L. Wuttisittikulkij and M. Alibakhshikenari, "Speech emotion recognition using convolution neural networks and multi-head convolutional transformer," *Sensors* **23**(13), 6212 (2023).
[24] P. Khan, P. Ranjan and S. Kumar, "AT2GRU: A human emotion recognition model with mitigated device heterogeneity," *IEEE Trans. Affect. Comput.* **14**(2), 1520–1532 (2023).
[25] P. Sripian, M. N. A. M. Anuardi, J. Yu and M. Sugaya, "The implementation and evaluation of individual preference in robot facial expression based on emotion estimation using biological signals," *Sensors* **21**(18), 6322 (2021).
[26] J. Wuth, P. Correa, T. R. Nuñez, M. Saavedra and N. B. Yoma, "The role of speech technology in user perception and context acquisition in HRI," *Int. J. Soc. Robot.* **13**(5), 949–968 (2021).
[27] A. Weiss and K. Spiel, "Robots beyond science fiction: Mutual learning in human-robot interaction on the way to participatory approaches," *AI Soc.*, 501–515 (2021).
[28] A. E. H. rsqb, R. Errattahi, F. Z. Salmam, T. Hain and H. Ouahmane, "Evaluation of the effectiveness and efficiency of state-of-the-art features and models for automatic speech recognition error detection," *J. Big Data* **8**(1), 1–16 (2021).
[29] M. H. Korayem, S. Azargoshasb, A. H. Korayem and S. Tabibian, "Design and implementation of the voice command recognition and the sound source localization system for human-robot interaction," *Robotica* **39**(10), 1779–1790 (2021).
[30] C. A. Monje, P. Pierro and C. Balaguer, "A new approach on human-robot collaboration with humanoid robot RH-2," *Robotica* **29**(6), 949–957 (2011).
[31] D. Fu, F. Abawi, H. Carneiro, M. Kerzel, Z. Chen, E. Strahl, X. Liu and S. Wermter, "A trained humanoid robot can perform human-like crossmodal social attention and conflict resolution," *Int. J. Soc. Robot.* **15**(8), 1325–1340 (2023). doi: 10.1007/s12369-023-00993-3.
[32] U. Maniscalco, P. Storniolo and A. Messina, "Bidirectional multi-modal signs of checking human-robot engagement and interaction," *Int. J. Soc. Robot.* **14**(5), 1295–1309 (2022).
[33] K. Tatarian, R. Stower, D. Rudaz, M. Chamoux, A. Kappas and M. Chetouani, "How does modality matter? Investigating the synthesis and effects of multi-modal robot behavior on social intelligence," *Int. J. Soc. Robot.* **14**(4), 891–911 (2022).
[34] S. Lee, D. K. Han and H. Ko, "Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification," *IEEE Access* **9**, 94557–94572 (2021).
[35] C. M. A. Ilyas, R. Nunes, M. Rehm, K. Nasrollahi and T. B. Moeslund, "Deep emotion recognition through upper body movements and facial expression," *Int. Conf. Comput. Vision* **5**, 669–679 (2021).
[36] C. Sendra-Balcells, V. M. Campello, J. Torrents-Barrena, et al., "Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five African countries," *Sci. Rep.* **13**, 2728 (2022).
[37] M. Xia, A. Field and Y. Tsvetkov. Demoting Racial Bias in Hate Speech Detection. **In:** *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media* (2020) pp. 7–14.
[38] A. A. Nichol, E. Bendavid, F. Mutenherwa, C. Patel and M. K. Cho, "Diverse experts' perspectives on ethical issues of using machine learning to predict HIV/AIDS risk in sub-Saharan Africa: A modified Delphi study," *BMJ Open* **11**(e052287), e052287 (2021). doi: 10.1136/BMJOPEN-2021-052287.
[39] O. Papakyriakopoulos and A. Xiang, "Considerations for ethical speech recognition datasets," **In:** *Proceedings of Sixteenth International Conference on Web Search Data Mini.*, 1287–1288 (2023). doi: 10.1145/3539597.3575793.

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press

[40] S. Shahriar, S. Allana, S. M. Hazratifard and R. Dara, "A survey of privacy risks and mitigation strategies in the artificial intelligence life cycle," *IEEE Access* **11**, 61829–61854 (2023).
[41] A. Mwogosi and S. M. Kibusi, "Critical success factors for EHR systems implementation in developing countries: A systematic review," *Glob. Knowl. Memory Commun.*, (2024). doi: 10.1108/gkmc-05-2024-0264.
[42] K. Chard and K. Bubendorfer, "High performance resource allocation strategies for computational economies," *IEEE Trans. Parall. Distrib.* **24**(1), 72–84 (2013).
[43] J. Choi, Z. Gu, J. Lee and I. Lee, "Impedance matching control between a human arm and a haptic joystick for long-term," *Robotica* **40**(6), 1880–1893 (2022).
[44] A. Romero, F. Bellas and R. J. Duro, "A perspective on lifelong open-ended learning autonomy for robotics through cognitive architectures," *Sensors* **23**(3), 1611 (2023).

Cite this article: I. Kishor, U. Mamodiya, S. Saini and B. Bossoufi (2025). "Voice-enabled human-robot interaction: adaptive self-learning systems for enhanced collaboration", Robotica **43**, 2143–2171. https://doi.org/10.1017/S0263574725000438

https://doi.org/10.1017/S0263574725000438 Published online by Cambridge University Press