MAIT Journal of Science & Technology
Issue No. 1, Volume No. 2, December 2025, Page No. 48-56
ISSN(Online)

Voice-Operated UAV System Using AI with Live Two directional Communication

Sharvi Aggarwal[1], Sneha Desai[2], Saarthak Kamra[3], Lalit Agarwal[4]

[1,2,4]Department of Electronics and Communication Engineering, Maharaja Agrasen Institute of Technology, New Delhi, India

[3]Department of Computer Science and Engineering, Maharaja Agrasen Institute of Technology, New Delhi, India

* Corresponding Author. E-mail: lalitagarwal@mait.ac.in

Abstract— Modern Drones are rapidly transforming into intelligent, interactive and smart aerial systems that are capable of communicating in real-time and make autonomous decisions. This paper investigates the mechanism behind voice command recognition, two-way communication, and virtual assistant integration to boost UAV capabilities. The application of Hidden Markov Models (HMM) and Maximum Likelihood Linear Regression (MLLR) techniques in the field of speech recognition allows drones to understand and process voice instructions. The two-way voice communication is facilitated through a WebRTC-based communication system using LTE, and thus drones can serve as flying communication stations. Furthermore, the use of a Raspberry Pi system helps delivers real-time telemetry, weather information, and flight diagnostics.

Keywords: Voice-controlled UAVs, two-way communication systems, virtual assistant implementation, WebRTC technology, Drone voice recognition, intelligent aerial systems.

I. INTRODUCTION

With the advancement of technology in drones, it has become easier for humans to communicate with them. Unlike using a manual controller or programmed route, a drone can now be controlled by voice command and also receive immediate feedbacks too. Using voice recognition technology, two-way communication, and artificial intelligence virtual assistant, the UAVs can easily be turned into interactive systems that can respond intelligently [1], [2].

The authors in [3], [4] discusses some of the technologies behind the same. HMM and MLLR speech recognition technologies can be used to recognize voice commands, making the control very convenient and hands-off. A WebRTC-based communication system that operates on LTE network makes the drones have real-time interaction capabilities using their microphone and speaker. In addition, a virtual assistant running on a Raspberry Pi computer provides live telemetry data, weather forecast, and other diagnostic services.

Not only does this combination improve usability, but it also makes operation easy in an environment that is quite complex. It enables the drones to be able to respond to the situation immediately without any human intervention. In addition, this makes the UAVs accessible to users, as they can control them even without any specialized knowledge.

II. RELATED WORK

According to Nair et al. [1] a voice-controlled quad-copter was implemented using basic keyword matching and RF based communication. Fayjie et al. [2] developed a voice-enabled smart drone control system with the help of deep learning. Despite achieving better recognition performance, the system could not include two-way communication functionality. Kalkan et al. [3] created a two-way communication system that was based on UAVs with the use of analog FM modulation. It suffered from poor noise immunity also. Kishore kumar et al. [4] proposed a drone personal assistant based on AI. However the system was completely dependent on cloud processing. Nand and Mathiyazaghan [5] built a UAV controlled through voice commands. It used a very simple threshold-based detection mechanism. Rahim et al. [6] introduced a VOC-Drone which was basically an AI-enabled voice controlled drone system but this system lacked speaker adaptation or two-way communication.

Table 1: Comparison with Existing Systems

Feature	[1]	[2]	[4]	[6]	Proposed
Voice Control	✓	✓	✓	✓	✓
HMM + MLLR	—	—	—	—	✓
Speaker Adapt.	—	—	—	—	✓
Two-Way Comm.	—	—	—	—	✓
Virtual Assistant	—	—	✓	—	✓
Telemetry	—	✓	—	✓	✓
NLP	—	—	—	—	✓

III. SPEECH RECOGNITION: HOW IT WORKS

In this section, we describe the main techniques used in our system. The voice recognition workflow has three major parts:

(i) Feature extraction using MFCCs,

(ii) Command classification using HMMs, and

(iii) Adaptation of the system to individual voices by the use of MLLR.

A. MFCC Feature Extraction

Before the system can recognize speech the raw audio signal must be converted into numerical data that the computer can easily work with and understand. This conversion or transformation can be achieved using Mel-Frequency Cepstral Coefficients (MFCCs) [7], [8].

1) Pre-emphasis: The input audio signal is passed through a filter to enhance the higher frequency components:

$$y(t) = x(t) - 0.97 \cdot x(t-1) \tag{1}$$

Where $y(t)$ is the filtered output signal, $x(t)$ is the current input sample, and $x(t-1)$ is the previous input sample. The coefficient 0.97 controls the degree of pre-emphasis applied.

2) Framing & Windowing: The audio signal is divided into overlapping frames of about 25 milliseconds for further analysis.

3) FFT & Mel Filterbank: Each segment is converted from time domain to frequency domain and passed through triangular filters on the Mel scale:

$$Mel(f) = 2595 \times \log_{10}(1 + f/700) \tag{2}$$

Where $Mel(f)$ is the perceived frequency on the Mel scale, and f is the corresponding physical frequency in Hz.

4) DCT: The final 13 MFCC coefficients is produced through The Discrete Cosine Transform.

B. Hidden Markov Model (HMM)

After the successful extraction of the MFCC features, Hidden Markov Models are used to match them to specific commands. An HMM is a statistical model or approach where speech is treated as a sequence of hidden states, and each producing observable features. The HMM may be defined or stated by transition probabilities (A), observation probabilities (B), and initial state probabilities (π) [7], [9].

The observation probability is modeled with the help of a Gaussian Mixture Model as:

$$b\_j(O\_t) = \Sigma\, c\_jm \cdot N(O\_t;\, \mu\_jm,\, \Sigma\_jm) \tag{3}$$

Where, $b\_j(O\_t)$ is the observation probability of feature vector O_t in state j, c_jm is the mixture weight for the m-th Gaussian component, μ_jm is the mean vector, and Σ_jm is the covariance matrix of that component.

During recognition, the Viterbi algorithm determines the most likely command. The model is trained and carried out using the Baum-Welch algorithm [9], [10].

C. MLLR Speaker Adaptation

MLLR solves and addresses the problem of different speakers who have unique and different voice characteristics. A linear transformation can be applied to the HMM parameters: [10], [11]

$$\hat{\mu} = A \cdot \mu + b \tag{4}$$

Where A is a transformation matrix and b is a bias vector, estimated from 3–4 sentences spoken during calibration. This is something that makes our system unique; the drone adapts to each pilot's voice.

IV. PROPOSED SYSTEM ARCHITECTURE

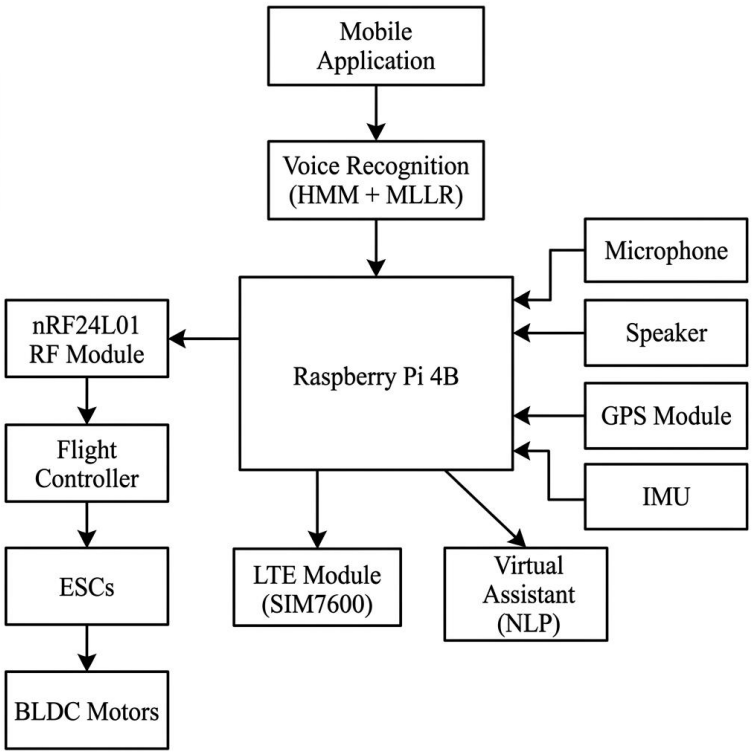

Fig. 1: Proposed UAV System Architecture

A. Voice Recognition Setup and Calibration

Before the flight takeoff, the pilot is required to speak 3–4 predefined sentences. The system then extracts MFCC features and the MLLR module creates a unique voice profile. This approach also ensures that only verified voice inputs are recognized by the system.

B. Command Flow and Signal Transmission

The pilot provides commands through the app. These commands are analyzed and verified using the combined HMM and MLLR-based recognition system. Once validated, the commands are sent or transmitted via nRF24L01 RF module to the flight controller. If in case any invalid command is detected, the drone automatically maintains its position by hovering and displays a warning notification in the application.

C. Pre-defined Voice Commands

The Command set includes instructions such as: "Throttle [value] percent,""Pitch forward [value],""Roll right [value],""Yaw right [value],""Hold position," and "Return home."

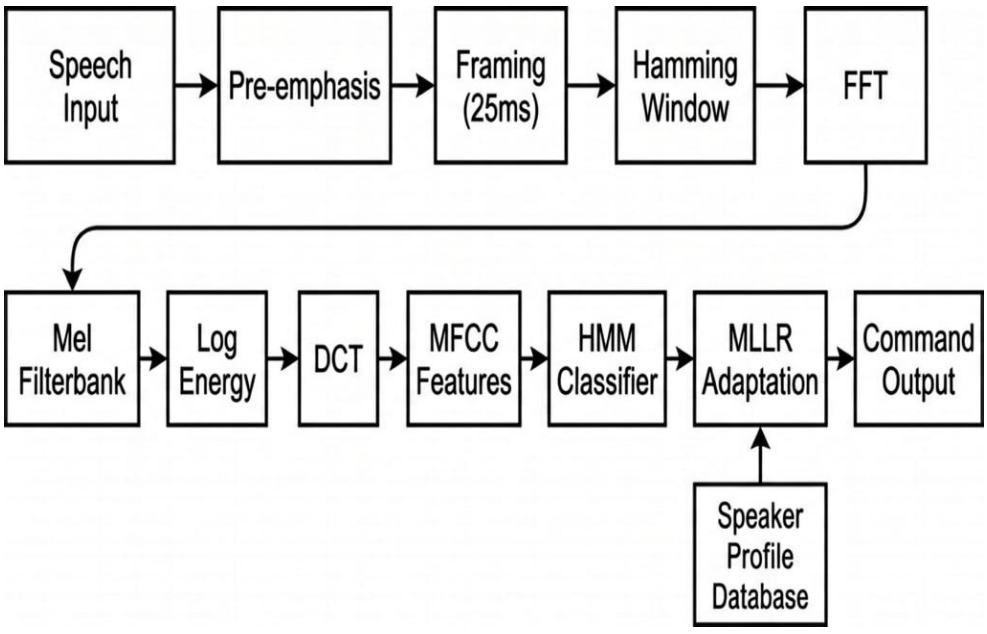

Fig. 2: Voice Recognition Pipeline

V. TWO-WAY COMMUNICATION FOR UAVS

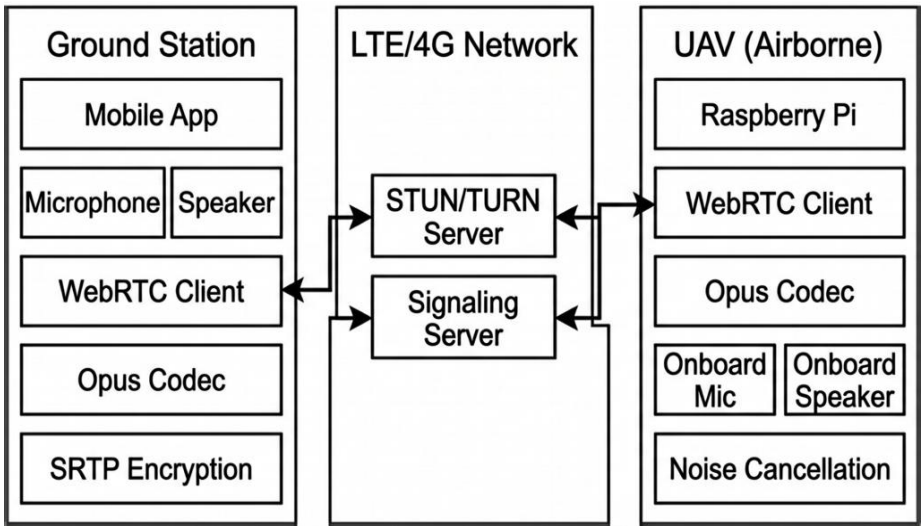

Fig. 3: Two-Way Communication Architecture

A. Activation and Audio Transmission

The communication system is operated by using voice command. A WebRTC based peer-to-peer session is established over LTE (SIM7600) connectivity. Also even during the Ground-to-Drone communication: audio is compressed (Opus), encrypted (SRTP), and transmitted via LTE. Drone-to-Ground: onboard mic captures audio and transmits back [12].

B. Noise Reduction and Session Management

Noise reduction algorithms are used to filter propeller noise (3–8 kHz) while preserving and not making any changes in the voice (300 Hz – 3.4 kHz). The user can deactivate communication by voice command to conserve power.

VI. VIRTUAL ASSISTANT INTEGRATION

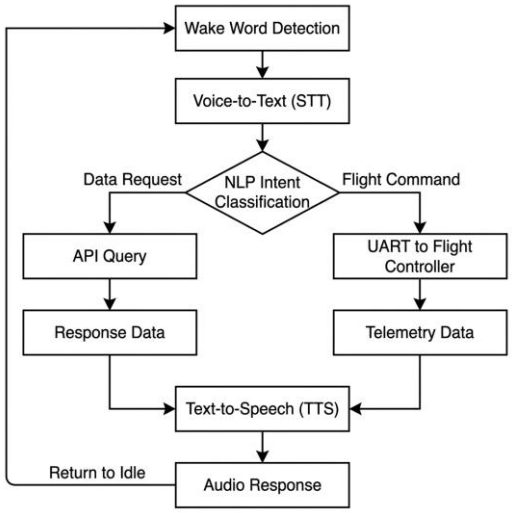

Fig. 4: Virtual Assistant Processing Flow

The virtual assistant is triggered using a predefined wake word. After activation the Spoken instructions are converted into text (DeepSpeech/Whisper) and analyzed by NLP (spaCy/NLTK). For data requests, it queries APIs. For flight specific operations, the assistant changes information with the flight controller via UART/MAVLink. Responses are then generated using text-to-speech synthesis [13-15].

VII. COMPONENTS REQUIRED

A. Hardware

• Raspberry Pi (Model 4B or later)
• Flight Controller (Pixhawk, Betaflight, or Ardupilot)
• Electronic Speed Controllers (ESCs)
• Brushless DC Motors
• RF Module (nRF24L01 series)
• LTE Module (SIM7600 or equivalent)
• Microphone and Speaker
• Lithium Polymer (LiPo) Battery
• IMU (Inertial Measurement Unit) and GPS Module

B. Software

• Raspbian OS, Python & C++
• HMM & MLLR Libraries
• Voice-to-Text Engine (DeepSpeech, Google Speech API, or Whisper)
• Text-to-Speech Engine (Festival TTS, eSpeak, or Google TTS)
• MAVLink Protocol and UART Communication
• NLP Module (spaCy, NLTK, or OpenAI-based models)
• WebRTC Framework.

VIII. EXPECTED PERFORMANCE

Based on existing research and published benchmarks in the field of embedded speech recognition and WebRTC communication, the expected performance of the proposed system is summarized below.

Table 2 shows the expected voice recognition accuracy under different noise conditions. The MLLR adaptation gain shows how much improvement we expect from calibrating the system to a specific speaker's voice.

Table 2: Expected Voice Recognition Performance

Metric	Clean	Moderate Noise	High Noise
Command Accuracy	97–99%	88–95%	70–85%
Response Latency	<200ms	<300ms	<500ms
MLLR Improvement	+2–3%	+5–8%	+8–12%

Table 3 shows the expected communication performance based on WebRTC-over-LTE benchmarks.

Table 3: Expected Communication Performance

Metric	Expected Value
End-to-End Latency	<100ms
Jitter	<30ms
Packet Delivery Rate	>99.9%
Audio Bandwidth	32–128 kbps

Table 4 shows the projected performance of the virtual assistant in handling speech recognition, query processing, and real-time responses.

Table 4: Expected Virtual Assistant Performance

Metric	Expected value
Wake Word Detection Accuracy	95–98%
Speech-to-Text Accuracy (Whisper/DeepSpeech)	90–96%
NLP Intent Classification Accuracy (spaCy/NLTK)	88-93%
Telemetry Query Response Time	<500ms
Weather API Fetch Time	1–3s (LTE dependent)
TTS Response Generation Time	<300ms
False Wake Word Activation Rate	<2%
CPU Usage on Raspberry Pi 4B	35-55%

IX. CONCLUSION AND FUTURE WORK

The incorporation of voice commands, virtual assistants and two-way communication into UAVs makes drones more interactive. Voice enabled control makes the operation simple while the virtual assistant provides real-time data. The combination of HMM and MLLR based speaker adaptation enables quick calibration for different users. Additionally, Two-way communication makes drones function or work as mobile communication platforms for security, search and rescue, and remote assistance. With the increasing progress of AI, IoT, and 5G, future drones may become fully autonomous, that will be capable for smarter and more responsive UAVs. Additionally the modular structure of the drone system ensures high scalability, making it possible to integrate features and allow future enhancements like gesture-based control, coordination among multiple drones, and edge AI inference. Ultimately, the fusion of embedded AI, real-time communication, and adaptive speech recognition enables voice-controlled UAVs to transform from simple tools into intelligent collaborative agents capable of applications in disaster management, surveillance, and precision logistics.

In future, the following points are to be addressed:

1. Enhanced voice recognition with noise filtering: Improving the HMM and MLLR-based voice recognition by integrating deep learning models to better differentiate the user's voice from background noise.

2. Customizable voice profiles: Allowing multiple users to train the system with their voices, enabling multi-user authentication and personalized drone operation.

3. Smart virtual assistant with context awareness: Enhancing the virtual assistant to understand multi-step commands and provide predictive insights based on real-time drone status.

4. AI-powered obstacle avoidance: Using liDAR sensors , ultrasonic sensors and computer vision techniques to independently sense and avoid obstacles.

5. Advanced communication features: Adoption of 5G technology to ensure reliable, instant and low latency communication.

6. Inclusion of Gesture and Multi modal control: providing gesture recognition in place of voice commands to increase flexibility.

7. Increase in flight duration: Implementing solutions like wireless charging , swappable batteries or solar powered systems.

REFERENCES

[1] N. M. Nair, P. Kale, K. Narayanan, and S. Salian, "Voice Controlled Quadcopter," International Journal ofResearch in Advent Technology, vol. 6, no. 3, March 2018.

[2] A. R. Fayjie, D. Oualid, A. Ramezani, and D. J. Lee, "Voice Enabled Smart Drone Control," Proc. IEEEInternational Conference on Consumer Electronics, 2019.

[3] Y. Kalkan, O. Avcı, T. Ulutaş, E. C. Akar, and B. Koksal, "Simple Design and Implementation of Two-WayCommunication System through UAV," Balkan Journal of Electrical & Computer Engineering, vol. 11, no. 1,January 2023.

[4] A. Kishorekumar, E. Ezhilarasan, and R. Parthiban, "Intelligent Drone based Personal Assistant using Artificial Intelligence (AI)," Int. Journal of Trend in Scientific Research and Development, vol. 2, Issue 3, Mar-Apr2018.

[5] S. S. Anand and R. Mathiyazaghan, "Design and Fabrication of Voice Controlled Unmanned Aerial Vehicle,"International Journal of Robotics and Automation (IJRA), Vol. 5, No. 3, September 2016.

[6] S. M. Rahim, A. Das, N. Kumar, S. Mukherjee, and R. Mishra, "VOC-Drone: AI Powered Voice ControlledAerial System," JETIR, Volume 11, Issue 12, December 2024.

[7] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989.

[8] Practical Cryptography, "Mel Frequency Cepstral Coefficient (MFCC) tutorial," Available:https://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[9] GeeksforGeeks, "Hidden Markov Model in Machine Learning," Available:https://www.geeksforgeeks.org/hidden-markov-model-in-machine-learning/

[10] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation ofContinuous Density Hidden Markov Models," Computer Speech & Language, vol. 9, pp. 171–185, 1995.

[11] S. Young et al., "The HTK Book (for HTK Version 3.4)," Cambridge University Engineering Department,2009.

[12] WebRTC Official Documentation, Available: https://webrtc.org/getting-started/overview

[13] spaCy NLP Library Documentation, Available: https://spacy.io/usage

[14] A. Hannun et al., "Deep Speech: Scaling up end-to-end speech recognition," arXiv preprint arXiv:1412.5567,2014.

[15] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak SupeSrvision," in Proc. ICML, 2023.