PROJECT REPORT

Offline Hindi Voice Assistant using Raspberry Pi 4:

1. Abstract:

This project presents the design and implementation of a fully offline Hindi voice assistant using the Raspberry Pi 4 Model B. The system integrates speech recognition, natural language processing, and speech synthesis without relying on internet connectivity.

The assistant uses Vosk for offline speech-to-text conversion and llama.cpp to run a lightweight Large Language Model locally. Responses are generated and converted to speech using eSpeak NG.

The system demonstrates a low-cost, privacy-preserving AI assistant suitable for rural and offline environments.

---

2. Introduction:

Voice assistants like Alexa and Google Assistant depend heavily on cloud infrastructure. This project aims to develop a fully offline alternative that runs on low-power hardware.

Key goals:

- No internet dependency
- Support for Hindi language
- Real-time interaction
- Low-cost implementation

3. System Architecture:

Block Diagram:

Microphone → Speech Recognition → LLM Processing → Text Output → Speech Synthesis → Speaker

Modules Used:

1. **Speech Input Module**
2. **Speech Recognition Engine**
3. **Language Model (LLM)**
4. **Response Generator**
5. **Speech Output Module**

4. Hardware Requirements:

Component	Description
Raspberry Pi 4 (4GB)	Main processing unit
USB Microphone	Audio input
3.5mm Jack / USB Speaker	Audio output
SD Card (16GB+)	OS and storage
Power Supply (5V/3A)	Stable power

5. Software Requirements:

- OS: Raspberry Pi OS (64-bit)
- Python 3.9+
- Libraries:
    - vosTools:
    - git
    - cmake
    - ffmpeg

6. Methodology:

Step 1: Audio Capture:

The microphone captures real-time voice input.

Step 2: Speech Recognition:

Audio is processed using the Vosk Hindi model to convert speech into text.

Step 3: Language Processing:

The recognized text is passed to a lightweight LLM using llama.cpp.

Step 4: Response Generation:

The LLM generates a contextual response.

Step 5: Speech Output:

The generated response is converted into Hindi speech using eSpeak.

7. Implementation Details:

Speech Recognition:

- Model: Vosk Hindi Small Model
- Sampling Rate: 16 kHz
- Output: JSON text

LLM Integration:

- Model: TinyLlama 1.1B (Quantized GGUF)
- Engine: llama.cpp
- Optimization: Q4 quantization for low memory usage

Speech Output:

- Engine: eSpeak NG
- Language: Hindi (hi)

8. Results:

The offline Hindi voice assistant was successfully implemented and tested on the Raspberry Pi 4 Model B. The voice assistant was able to complete an end-to-end voice interaction without the need for any internet connection.

The voice assistant was able to successfully translate speech inputs into text using the speech recognition module and provide a meaningful response using the local language model. The response was then converted into speech.

Performance Metrics:

- Response Time: 5–15 seconds
- RAM Usage: ~2.5 GB
- CPU Usage: High (80–100%)

9. Hardware Utilization:

Resource	Usage
CPU	High (LLM inference)
RAM	Moderate (2–3 GB)
Storage	~3–5 GB
Audio I/O	Active

10. Optimization Techniques:

- Use of **quantized models (Q4)**
- Reduced context size (n_ctx)
- Limited token generation
- Efficient C++ backend (llama.cpp)
- Lightweight Vosk model

11. Advantages:

- Fully offline operation
- Privacy-preserving
- Low-cost system
- Works in low connectivity areas

12. Limitations:

- Slower response time
- Limited model intelligence compared to cloud AI
- High CPU usage

13. Applications:

- Rural AI assistants
- Smart farming (AgriVision integration)
- Education tools
- Offline automation systems

14. Future Enhancements:

- Wake word detection ("सुनो")
- GPIO control for IoT devices
- Better Hindi TTS models
- Hardware acceleration (NPU/Coral USB)
- Improved LLM models

15. Conclusion:

The project successfully demonstrates a fully offline AI voice assistant using Raspberry Pi. It proves that intelligent systems can be deployed without internet dependency while maintaining functionality and privacy.

16. References:

1. Vosk Documentation
2. llama.cpp GitHub
3. eSpeak NG
4. Raspberry Pi Official Documentation