Voice Authentication Access Control System

Zexi Lu Zhikun Peng Zhengyuan Yao

***Abstract*—This project presents a multi-user voice authentication system integrating a Raspberry Pi 4B and Sony Spresense development board. The system utilizes Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction and Gaussian Mixture Models (GMM) for speaker verification. The Raspberry Pi handles the computationally intensive tasks of MFCC extraction and GMM execution, while the Spresense board manages audio recording and LED control for user feedback. This division of labor optimizes performance and power efficiency. The system's real-time, low-power voice recognition capability makes it suitable for smart home and IoT applications. Key innovations include the integration of advanced voice recognition techniques in an embedded system and the efficient use of resources between the two hardware components. The project demonstrates the potential for enhancing security and user experience in various IoT scenarios, with particular emphasis on smart home environments. Experimental results show promising accuracy in distinguishing between registered and unregistered users, highlighting the system's effectiveness in multi-user authentication scenarios.**

***Index terms*—Sony Spresense, Machine Learning, Voice Recognition**

I. Introduction

Voice authentication has emerged as a promising solution to the growing need for secure and user-friendly authentication methods in smart homes and IoT devices. This project aims to develop a real-time, energy-efficient voice authentication system capable of managing multiple users.

The system's operation begins when a user presses a button, triggering a 5-second audio recording on the Sony Spresense board. This audio data is then transmitted to a Raspberry Pi via serial connection for processing. The Raspberry Pi extracts MFCC features and performs GMM-based speaker verification. The verification outcome is communicated back to the Sony Spresense board, which provides visual feedback through LED indicators: a continuous green light for successful verification, or a flashing red light for failed attempts.

This approach effectively addresses key challenges in voice authentication, including real-time performance, energy efficiency, and multi-user support, making it particularly well-suited for smart home applications. By successfully integrating advanced voice recognition technology into resource-constrained embedded systems, this project demonstrates the potential for enhancing both user convenience and security in everyday environments.

The subsequent sections of this report will provide a comprehensive overview of existing voice recognition systems, detail the system's architecture and implementation, highlight innovative aspects, discuss the utilization of Spresense functionality, present experimental results, and explore potential improvements and future applications.

II. Project Technical Description

In this section we would show you the system structure and the processing procedure.

A. *Hardware Configuration*

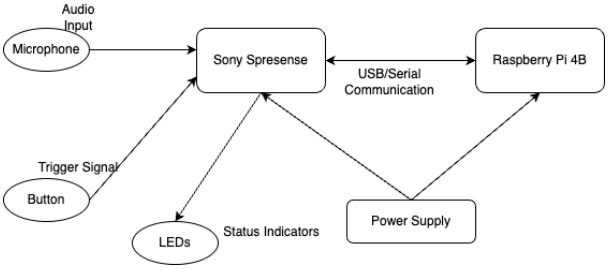

Figure 1: Hardware component

1. Raspberry Pi 4B:

At first we planned to use Sony Spresense to handle the whole audio processing, including the voice feature extraction(MFCC calculation) and similarity calculation(GMM scoring). But we spent a lot of effort, only found the Spresense board might have some performance limit running this task. Hence we decided to use Raspberry Pi to do the recognition part. This micro computer served as cental processing unit, responsible for:

- Receiving and processing the audio information provided by Sony Spresense board.
- Sending recognition status back to Sony Spresense using GPIO.

2. Sony Spresense: As a ultra-low power micro controller, it features:

    - Multiple high performance Cortex-M4F cores collaboration.
    - Massive RAM support.

- Hi-Fidelity audio processing provided by its high performance ADC.
- Low power consumption

In this project, it plays a vital role by providing high quality audio data and superior stable quality.

Pin Control: Spresense manages the LED lights' states based on verification results, providing immediate visual feedback.

3. Other Components:

Button: A physical button is used to initiate the recording process by sending a high-level signal to the Sony Spresense board.

LED Lights: Three different colored LEDs (blue, green, and red) are used to indicate the system's state (recording, verification success, and verification failure).

Independent Power Supply: An power bank is utilized to ensure a stable and adequate power source for both the Raspberry Pi and the Sony Spresense board.

B. *Software Design*

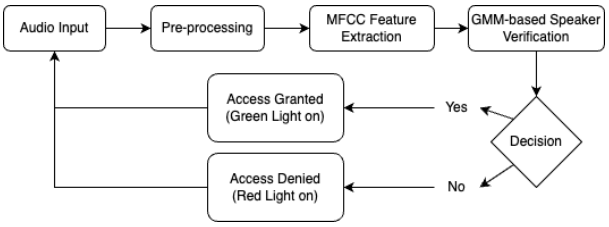

Figure 2: Authentication Process

1. MFCC Feature Extraction: Data preprocessing and feature extraction, simulating how human hears voices and making the audio more readable for machine learning models

2. GMM Training and Execution: The machine learning models that can calculate similarity between input and model data. It is a mature, low performance requirement and high accuracy model. In this project it is used in the following stages:

   - Training: GMMs are trained using MFCC features from a dataset of voice recordings. The Expectation-Maximization (EM) algorithm estimates the parameters, including means, covariances, and mixture weights.
   - Execution: For new audio inputs, MFCC features are extracted, and the likelihood of the features belonging to each speaker's GMM is calculated. The speaker with the highest likelihood is identified.

3. Data Communication:

Serial Interface: The Raspberry Pi 4B and Sony Spresense board communicate via a serial interface. Audio data is transmitted from the Spresense to the Raspberry Pi, and verification results are sent back to the Spresense.

4. LED Control:

Verification Feedback: Based on the verification results, the Sony Spresense board controls the LEDs. A green LED indicates successful verification, while a red LED indicates failure.

C. *Implementation Details*

1. Training of the model:

Model training and thresholds generation were performed on a laptop computer using Python, with a total of six audio recordings of the phrase "The quick fox jumps nightly above the wizard".

2. Audio files:

The all of audios are recorded in WAV format at a sampling rate of 48,000 Hz, with 16-bit depth in mono channel for a duration of 5 seconds, resulting in a total size of 480 KB per recording.

3. Speaker Verification:

The Raspberry Pi will continuously monitor whether the Spresense's USB MSC (Mass Storage Class) is turned on, and when it detects that it is on, it will automatically load the audio files stored on the SD card that need to be verified.

The probability is obtained by comparing the audio with the existing model after MFCC and GMM processing, and if it exceeds the corresponding threshold, the validation will pass, and vice versa, it will fail.

By implementing these technical components and processes, our project ensures a robust and efficient voice authentication system suitable for multi-user environments in smart home applications.

III. Novelty of the Proposed Application

A. *Integration of Advanced Voice Recognition Techniques in Embedded Systems*

1. Efficient Use of Resources:

Division of Labor: One of the most notable innovations is the efficient distribution of tasks between the Raspberry Pi 4B and the Sony Spresense board. The Raspberry Pi 4B handles the computationally intensive processes of MFCC feature extraction and GMM-based speaker verification, while the Sony Spresense board is dedicated to audio recording and pin control. This division allows the system to leverage the

strengths of each component, ensuring high performance without overburdening a single device.

Low Power Consumption: The Sony Spresense board is known for its low power consumption, making it an ideal choice for continuous audio recording and control tasks. This characteristic is crucial for embedded systems that need to operate efficiently over extended periods.

2. Real-Time Performance:

Minimized Latency: By offloading the intensive computation to the Raspberry Pi 4B and using a high-speed serial interface for data transmission, the system minimizes latency, ensuring near real-time response. This real-time capability is essential for applications such as security systems and smart home automation, where prompt action is required.

B. Multi-User Voice Authentication

1. Robust Feature Extraction and Modeling:

MFCC for Feature Extraction: The use of Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction provides a robust representation of the audio signal, capturing essential characteristics that differentiate between speakers. MFCC is a proven technique in speech processing, offering high accuracy in capturing phonetic details.

GMM for Speaker Verification: Gaussian Mixture Models (GMM) are employed for speaker verification, modeling the probability distribution of the extracted MFCC features. GMMs are particularly effective in handling the variability in speech patterns, enhancing the system's accuracy and reliability.

IV. Spresense kit Usage

The Sony Spresense development board plays a crucial role in the proposed voice authentication system, leveraging its advanced capabilities to handle audio recording and peripheral control with high efficiency and low power consumption. This section details the specific usage of the Spresense kit in our project, emphasizing its contributions to the overall system architecture and functionality.

A. Audio Recording

1. High-Quality Audio Capture:

Advanced ADC: The Spresense board is equipped with a high-performance Analog-to-Digital Converter (ADC), which ensures high-fidelity audio recording. This feature is critical for capturing the nuances of voice signals necessary for accurate feature extraction and speaker verification.

2. Real-Time Audio Processing:

Low Latency: The Spresense board processes audio in real-time, providing low-latency performance essential for immediate feedback in voice-controlled applications. This capability is particularly important for the authentication system, where timely responses are crucial for user experience.

Noise Reduction: Built-in audio processing features, such as noise reduction and normalization, enhance the quality of the recorded audio. These preprocessing steps ensure that the subsequent feature extraction and modeling stages operate on clean and reliable audio data.

B. Data Transmission to Raspberry Pi

Reliable Interface: The Spresense board uses a serial interface to transmit the recorded audio data to the Raspberry Pi 4B. This method ensures reliable and efficient data transfer between the two devices, enabling seamless integration of their respective functionalities.

Synchronization: The serial communication is synchronized to ensure that audio data is transmitted and received without loss or delay. This synchronization is vital for maintaining the integrity of the audio signal during the feature extraction and modeling processes on the Raspberry Pi.

C. Peripheral Control

1. LED Control:

Utilizing its GPIO pins, the Spresense board manages the states of the LED lights. When the verification is successful, the green LED lights up continuously. If the verification fails, the red LED flashes, providing a clear and intuitive indication of the system's status.

2. Button Input:

The Spresense board also handles button input through its GPIO pins. When a user presses the button, it initiates the entire verification process. This action triggers the board to start recording the user's voice for the subsequent authentication procedure, making the button an essential component for starting the voice verification sequence.

V. Test Results & Future Expectations

A. Test Result

To evaluate the performance of the system, we conducted tests with three users. Users 1 and 2 had trained models, while User 3 did not. Each user conducted 10 tests. The results are summarized below:

User	Test Number	Success Rate	Failure Rate
User 1	10	60%	40%
User 2	10	70%	30%
User 3	10	0%	100%

Table 1: Testing results for multi-user voice authentication

It can be observed from that the accuracy of successful recognition of registered users is not very high, but strange users who are not registered can be recognized effectively.

B. Practical Application in Smart Home Environments

1. Voice-Controlled Device Operation:

   LED Control: The system demonstrates a practical application of voice authentication by controlling LED lights based on verification results. This functionality can be extended to other smart home devices, showcasing the versatility and potential of voice-based control.

2. Enhanced Security and Convenience:

   Voice recognition provides enhanced security over traditional methods like passwords or Pins and offers a non-intrusive, user-friendly interaction, ideal for smart home applications.

C. Potential for Broader Applications

1. Internet of Things (IoT):

   Integration with IoT Devices: The architecture of the system allows for easy integration with various IoT devices, enabling voice-controlled operations across a range of applications. This integration highlights the potential for expanding the system's use beyond smart homes to broader IoT ecosystems.

   Automation and Control: The ability to authenticate users and control devices based on voice commands can be extended to industrial automation, healthcare, and other sectors where hands-free operation and secure access are essential.

2. Adaptive and Scalable Design:

   Future Enhancements: The modular design of the system allows for future enhancements and upgrades, such as incorporating more advanced machine learning models or expanding the range of controlled devices. This flexibility ensures that the system can adapt to evolving technological advancements and user needs.

In summary, the proposed voice authentication system stands out for its efficient use of resources, real-time performance, multi-user capability, and practical application in smart home environments. Its innovative integration of advanced voice recognition techniques within an embedded system framework demonstrates significant potential for enhancing security, convenience, and user experience across various applications.

VI. Authors

First Zexi Lu is borned in Foshan, Guangdong, China on June 5, 2001. He is currently a Year 2 computer science and electrical engineering (DMT Pathway) undergraduate student in University of Liverpool, studying Xi'an Jiaotong-Liverpool University's 2+2 program. His main field of study are computer science and electrical engineering.

Second Zhikun Peng was born in Hefei, Anhui, China on September 29, 2002. He is a second-year undergraduate student at the University of Liverpool, United Kingdom, pursuing a Bachelor of Engineering (BEng) degree in Computer Science & Electronic Engineering. His primary research interests include embedded development and machine learning.

Third Zhengyuan Yao was born in Inner Mongolia, China on November 16, 2002. He is an undergraduate student in his second year at the University of Liverpool, United Kingdom. He is pursuing a Bachelor of Engineering (BEng) degree in Computer Science & Electronic Engineering. His major fields of study are embedded development and machine learning.