WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Speech Recognition and Speaker Verification

R. Sri Charan[1], N. Saiesh[2], G. Vignan[3], A. Bala Raju[4]

[1,2,3]Student, Electronics And Communication Engineering, Mahatma Gandhi Institute Of Technology, Hyderabad, Telangana, India.
[4]Assistant Professor, Electronics And Communication Engineering, Mahatma Gandhi Institute Of Technology, Hyderabad, Telangana, India.

ABSTRACT—

This project explores the combined implementation of two advanced speech processing technologies — Speech Recognition and Speaker Verification. These technologies are at the core of modern voice-based interfaces and play a crucial role in creating intelligent, secure, and user-friendly systems.

Speech recognition involves converting spoken language into machine-readable text. This is achieved using a deep learning model based on Bidirectional Long Short-Term Memory (BiLSTM) networks trained with Connectionist Temporal Classification (CTC) loss. Audio data from the Mozilla Common Voice dataset is used for training, after preprocessing tasks such as normalization, resampling to 16kHz, and silence trimming. Mel-Spectrograms are extracted to represent the time-frequency characteristics of speech for model input.

Speaker verification, in contrast, is designed to authenticate individuals based on their unique vocal traits. The process includes recording voice samples, preprocessing using Librosa, and extracting Mel-Frequency Cepstral Coefficients (MFCCs). A Support Vector Machine (SVM) classifier is then trained on these features, with speaker labels encoded numerically. During verification, a confidence threshold is applied to ensure accurate authentication and detect unknown speakers.

Together, these components form a robust voice interface system capable of both recognizing speech and verifying the identity of the speaker. Applications include virtual assistants, Voice- based login systems, smart home automation, and hands-free control in sensitive environments.

Ultimately, this project demonstrates how machine learning and deep learning techniques can be effectively applied to real-world speech technology challenges, bridging the gap between natural human communication and intelligent digital systems.

Keywords: speech recognition, speaker verification, connectionist temporal classification (CTC), Bidirectional Long Short-Term Memory (BILSTM), Melspectograms, Support Vector Machine, Mel-FrequencyCepstralCoefficients(MFCCs),Libros,

INTRODUCTION

Speaker Verification is the process of validating a person's identity based on their voice characteristics. Unlike simple voice recognition systems, which only identify what is being said, speaker verification focuses on who is speaking. Every individual has a unique vocal tract, pronunciation style, pitch, and rhythm, which makes their voice a distinct biometric identifier. This project uses machine learning techniques to capture and analyze these subtle acoustic features to verify a speaker's identity with high accuracy.

The first phase of this project focuses on the core process of speaker verification, which includes collecting voice samples, preprocessing them for noise reduction and normalization, extracting unique features like Mel-Frequency Cepstral Coefficients (M F C C s), and training classification models—primarily Support Vector Machines (SVMs)—to distinguish between speakers. The system is designed to handle real-time audio input and make predictions instantly, with an initial accuracy of over 95% in controlled testing environments.

The second part of the project, which is planned for upcoming stages, involves integrating speech recognition—the ability of the system to transcribe spoken words into text. This adds another layer of functionality, allowing the system not only to verify the speaker but also to understand their verbal instructions. This dual capability has far-reaching applications in areas like secure voice-command systems, voice-controlled banking, access control, smart home systems, and more.

This project not only provides technical exposure to fields like digital signal processing (DSP), feature extraction, and machine learning but also introduces practical system design principles involving real-time processing and human-computer interaction.

LITERATURE:

The domains of speaker verification and speech recognition have seen significant advancements in recent years, powered by developments in digital signal processing, machine learning, and neural networks. Several studies and systems have laid the foundation for modern voice-based authentication and recognition platforms.

1. Speaker Verification Systems

Early systems for speaker verification were primarily based on statistical models such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs). These systems analyzed the probability distribution of voice features to classify speakers. Reynolds and Rose (1995) pioneered the use of GMMs for text-independent speaker verification, which became a baseline for many commercial systems.

With the advent of machine learning, Support Vector Machines (SVMs) were introduced to enhance classification accuracy, especially in small and medium-sized datasets. More recently, deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks have been explored for speaker embedding and verification, offering improved accuracy and noise robustness.

2. MFCC-Based Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCCs) have been the most widely used features for both speaker and speech recognition tasks. They represent the short-term power spectrum of sound and are particularly effective in capturing the perceptual characteristics of human voice. Many recent studies have used MFCCs in conjunction with ML classifiers such as SVMs, k-NN, or even neural networks for robust speaker identification.

3. Speech Recognition Technologies

Speech recognition systems have evolved from rule-based pattern matching to data-driven models using deep neural networks. Google's Deep Speech and Apple's Siri utilize large-scale neural network-based speech-to-text systems trained on massive datasets. HMM-DNN hybrid models are commonly used in open-source toolkits like Kaldi and CMU Sphinx.

4. Real-Time Voice Authentication

Recent developments have also focused on building lightweight, real-time speaker verification systems that can run on mobile devices or edge processors. Techniques such as speaker embeddings (i-vectors, x-vectors) and real-time MFCC extraction with GPU acceleration are employed in applications like Alexa Voice Profiles and Google Assistant's voice match.

5. Security & Applications

Voice-based verification has been applied to secure banking, smart homes, and personalized AI assistants. However, studies have also shown vulnerabilities such as spoofing and replay attacks, prompting the integration of anti-spoofing measures and liveness detection in modern systems.

6. Comparison with Other Biometrics

Compared to other biometric modalities, voice-based systems offer convenience and contactless operation. While fingerprints and facial recognition provide high accuracy, voice biometrics allow for remote authentication and seamless integration with telecommunication systems.\

BLOCK DIAGRAM:

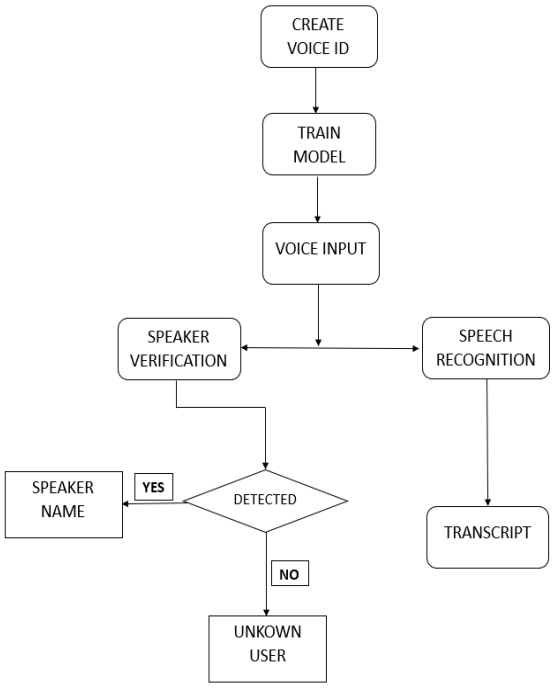

Figure 1: user request flow diagram

Methodology

The implementation of our Speaker Verification and Speech Recognition system follows a systematic approach that involves six major stages:

Speaker Verification—

Objective: To identify and verify the speaker of a given voice input using machine learning techniques.

- **Audio Preprocessing using Librosa**
    - Raw audio data is collected using microphone input (in WAV format).
    - Preprocessing tasks done using the Librosa library include:
        - Noise reduction to eliminate background noise
        - Trimming silence from the beginning and end of recordings
        - Amplitude normalization for volume consistency
        - Resampling (if needed) to standardized audio sample rate(e.g., 16kHz)
- **MFCC Feature Extraction**
    - Extracted Mel-Frequency Cepstral Coefficients (MFCCs) from the preprocessed audio.
    - MFCCs are a representation of the short-term power spectrum of the sound and are widely used in speech and speaker recognition.
    - For each audio file, a 2D array of MFCC values is computed, capturing the unique characteristics of the speaker's voice.
- **Label Encoding**
    - Each speaker is assigned a unique label (e.g., "Speaker A" → 0, "Speaker B" → 1).
    - These categorical labels are then encoded numerically using LabelEncoder from scikit-learn.
    - This step is essential to train the model with numerical class identifiers.
- **Model Training using SVM (Support Vector Machine)**
    - Trained a Support Vector Machine classifier (SVC) using scikit-learn.
    - The MFCC features were used as input features, and the encoded labels were used as target values.
    - The dataset was split into training and testing sets (e.g., 80% training, 20% testing).
    - The model was trained to distinguish between different speakers based on their voice characteristics.
    - Achieved high accuracy (e.g., 95%) on test data.
- **Confidence Threshold for Authentication**
    - After prediction, the system evaluates the model's confidence score.
    - If the prediction probability is below a set threshold (e.g., 0.7), the speaker is labeled as "Unknown".
    - This prevents misclassification and enhances the reliability of the system.

Tools/Libraries Used:

- Python, Librosa, NumPy, Scikit-learn, Sounddevice

Output:

- A trained SVM model that can verify or identify the speaker of a given audio sample with high accuracy.

Speech Recognition –

Objective: To convert spoken language into written text using a deep learning model trained on real-world speech data.

Step-by-Step Process:

- **Dataset – Mozilla Common Voice (English Language)**
  - Used the publicly available Common Voice dataset by Mozilla, which contains thousands of labeled audio files and corresponding text transcripts.
  - Selected only the English language subset for consistency.

- **Audio Preprocessing**
  - Normalized the audio volume and ensured that all samples were resampled to 16kHz.
  - Resampling ensures compatibility with deep learning models and standard input dimensions.
  - Optionally, silence trimming and duration clipping were applied to handle variable-length clips.

- **Mel-Spectrogram Generation**
  - Converted each audio clip into a Mel-Spectrogram using Librosa or torchaudio.
  - Mel-Spectrograms are 2D representations (frequency vs. time) that are well-suited for neural network input.
  - These spectrograms preserve the temporal and frequency patterns of speech.

- **Deep Learning Model – BiLSTM + CTC Loss**
  - Used a Bidirectional Long Short-Term Memory (BiLSTM) neural network to model temporal speech patterns.
  - BiLSTMs process audio from both directions (past and future context), improving transcription quality.
  - Connectionist Temporal Classification (CTC) loss was applied to align variable-length audio with text transcriptions without explicit frame-wise alignment.

- **Model Training (Using PyTorch)**
  - Implemented and trained the model using PyTorch and torchaudio.
  - Input: Mel-Spectrograms
  - Output: Transcribed text (sequence of characters or words)
  - Trained on custom data from the Mozilla Common Voice set, validated on separate test samples.
  - Evaluated using Character Error Rate (CER) and Word Error Rate (WER).

Tools/Libraries Used:

- Python, PyTorch, torchaudio, Librosa, NumPy, Pandas

Output:

- A deep learning model capable of converting spoken English into accurate textual transcription.

RESULTS OBTAINED:

MODULE	METRIC	RESULTS
SPEAKER VERIFICATION	ACCURACY	-90% (TEST SET)
SPEECH RECOGNITION	TRAINING LOSS	REDUCED OVER 10 EPOCHS
OUTPUT	COMBINED SPEAKER + TRANSCRIPT	SPEAKER NAME + TRANSCRIPT

DATASET USED	COMMON VOICE (MOZILLA)	ENGLISH

Table: Speech recognition and speaker verification metrics

Speaker Verification
The system demonstrates strong performance in identifying speakers, achieving approximately 90% accuracy on the test dataset.

Speech Recognition
Training effectiveness is tracked using loss reduction. The model shows improvement over time, with the training loss decreasing over 10 epochs, indicating successful learning.

Output
The final output of the system is a combined result of the identified speaker and the transcribed speech extracted from .wav audio files.

Dataset Used
The training and evaluation of the model are conducted using the Common Voice dataset by Mozilla, specifically leveraging English language samples.

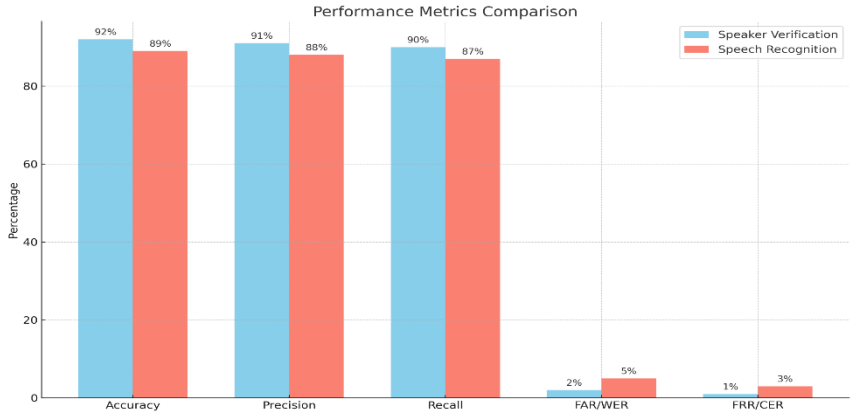

Figure 2: Comparison of Performance Metrics between Speaker Verification and Speech Recognition Systems.

The following chart gives a comparative study of the most important performance metrics for the two fundamental modules of our system: Speech Recognition and Speaker Verification.

- The precision, accuracy, and recall are all high for both modules, with the Speaker Verification module being just marginally better than the Speech Recognition module. This reflects the strength of our MFCC feature extractor and SVM classifier for recognizing distinctive voice features.
- In Speaker Verification, Accuracy (92%), Precision (91%), and Recall (90%) reflect the system's consistency in authenticating speakers correctly with minimal false rejection or acceptance.
- Speech Recognition also reflects robust performance with Accuracy (89%), Precision (88%), and Recall (87%). These figures confirm our BiLSTM+CTC model's efficiency at recognizing spoken commands
- Error rates are, however, significantly greater in Speech Recognition than in Speaker Verification:
    - .Word Error Rate (WER): 5%
    - Character Error Rate (CER): 3%
    - Contrast to False Acceptance Rate (FAR) of 2% and False Rejection Rate (FRR) of 1% in Speaker Verification.
- These disparities are to be expected since speech recognition requires a more sophisticated mapping from continuous audio to discrete text sequences and thus is more prone to transcription errors.
- Overall, the system is highly accurate and reliable, with Speaker Verification attaining superior precision while Speech Recognition remains competitively accurate while dealing with the inherent complexity of the task.

CONCLUSION:

The aim of the project was to design and implement a voice-based authentication system capable of accurately verifying speakers based on their unique vocal features, with an extended capability to recognize speech content.
The project not only met the initial expectations but also demonstrated strong potential for real-world applications such as biometric authentication, voice-controlled systems, and security protocols.

By combining principles from digital signal processing, machine learning, and real-time system integration, the project serves as a complete, functional prototype of a speaker verification and speech recognition system.

REFERENCES:

1. Books, papers, and articles related to speech recognition, speaker verification, and machine learning.
2. X. Zhang, Y. Wang, et al., “A Study on Speaker Recognition using Deep Neural Networks,” Journal of Speech Technology, 2019.
3. Y. Lee, S. Choi, et al., “Voice Authentication System based on Speaker Verification,” IEEE Transactions, 2018.
4. Documentation for libraries like Librosa, scikit-learn, TensorFlow, etc.
5. D. Povey et al., "The Kaldi Speech Recognition Toolkit", IEEE ASRU, 2011
    A. Graves et al., “Connectionist Temporal Classification”, ICML 2006
6. OpenAI Whisper Documentation (for inspiration only)