Speaker Recognition

A Project Report Submitted by

Abhishek Sahu & Chandra Mohan Singh Negi

in partial fulfillment of the requirements for the award of the degree of

M.Tech. in AI

Indian Institute of Technology Jodhpur

Computer Science Engineering

February, 2025

Declaration

I hereby declare that the work presented in this Project Report titled Speaker Recognition submitted to the Indian Institute of Technology Jodhpur in partial fulfillment of the requirements for the award of the degree of M.Tech. in AI, is a bonafide record of the research work carried out under the supervision of Richa Singh. The contents of this Project Report in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

Signature

Abhishek Sahu & Chandra Mohan Singh Negi

M23CSA504 & M23CSA512

Certificate

This is to certify that the Project Report titled Speaker Recognition Report, submitted by Abhishek Sahu & Chandra Mohan Singh Negi(M23CSA504 & M23CSA512) to the Indian Institute of Technology Jodhpur for the award of the degree of M.Tech. in AI, is a bonafide record of the research work done by him under my supervision. To the best of my knowledge, the contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Signature

Richa Singh

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor...

Abstract

The M.Sc./M.Sc.-M.Tech./M.Tech. Program of study requires each student to undertake research in the chosen area of study and to submit a thesis on it in consultation with the faculty member(s) supervising the same. The M.Sc./M.Sc.-M.Tech./M.Tech. Project is included in the curriculum with a view to synthesize the various components of the research work undertaken during the of the M.Sc./M.Sc.-M.Tech./M.Tech. Program at IIT Jodhpur. Creating a Project Report document of the research undertaken is part of the skill building training of the student in technical communications. Here, the emphasis is on presenting a technical matter in an objective written form.

This document is a record of the mandatory guidelines to be followed while preparing the of the Project Report document to be submitted at the end of the M.Sc./M.Sc.-M.Tech./M.Tech. Program. It prescribes typical contents that an M.Sc./M.Sc.-M.Tech./M.Tech. Project Report document usually should contain, and provides the format of its presentation. While most of these guidelines are prescriptive, some are subjective; but towards ensuring a relatively uniform style of presentation of all M.Sc./M.Sc.-M.Tech./M.Tech. Project Report being submitted at the Institute, these subjective guidelines are expected to help in setting at least a reasonable minimum expectation of the presentation level of the work accomplished in the research program.

All students pursuing M.Sc./M.Sc.-M.Tech./M.Tech. Program are urged to read the contents and form of this document carefully, and prepare their Project Report document as prescribed. It is hoped that this document will lead to a modest beginning at the Institute towards imparting education in professional written presentations.

Contents

List of Figures

List of Tables

Speaker Recognition

1 Objective

Analyze and evaluate modern speaker identification models, their strengths, limitations, and future potential.GitHub

1.1 What is Speaker Identification?

Speaker identification is the task of determining who is speaking based on an audio sample. In speech processing, speaker identification refers to identifying which person among a set of known speakers is speaking. This is distinct from speaker verification, where the task is to verify if the speaker is who they claim to be.

1.2 Real-World Importance

Speaker identification plays a significant role in several real-world applications, such as:

- **Security**: It can be used for authentication in banking, call centres, and secure access systems (e.g., voice biometrics).
- **Forensics**: Speaker identification helps identify speakers in criminal investigations or court cases involving audio evidence.
- **Speech Analytics**: : In call centres, speaker identification can help categorize conversations by different agents and customers.
- **Human-Computer Interaction**: Voice assistants (like Alexa, Siri, etc.) can identify individual speakers to provide personalized responses.

2 State of the Art (SOTA) Models

2.1 Methods Used in Speaker Identification

There are several approaches in speaker identification, ranging from classical methods to advanced deep learning-based techniques. The key methods are:

2.1.1 Classical Methods

- **Gaussian Mixture Models (GMM)**: Historically, GMMs have been used to model the acoustic features of a speaker's voice. These models use a probabilistic approach to describe the speaker's voice characteristics.

- **Hidden Markov Models (HMM)**: Often used in speech recognition and speaker identification tasks, HMMs model sequential data and are effective in capturing the temporal structure of speech.
- **i-vector / x-vector**: These methods have become standard in recent years for speaker identification tasks. i-vectors map the speech features into a fixed-size vector representation, and x-vectors represent more advanced versions of i-vectors with better discrimination power.

2.1.2 Deep Learning-Based Models

- **Convolutional Neural Networks (CNNs):**: CNNs have been used to learn speaker-specific features from spectrogram representations of the audio signal.
- **Recurrent Neural Networks (RNNs):**:): RNNs, especially Long Short-Term Memory networks (LSTMs), can capture temporal dependencies in speech, which is critical for modelling speech dynamics.
- **ResNet for Audio Classification:** : Using residual networks (ResNets) with spectrogram or Mel-spectrogram features is a recent trend in speaker identification.
- **Transformer-based Models:**: Recent work on transformer architectures (like Wav2Vec2) has demonstrated success in speaker recognition by using large amounts of unannotated data.

2.1.3 SOTA Frameworks

- **x-vector Embeddings**: These embeddings are derived from a deep neural network trained to classify speakers. They have outperformed traditional methods like GMM-HMM systems.
- **Deep Speaker Embedding:**: A Deep learning-based models like ECAPA-TDNN and ResNet-based architectures are also gaining prominence in speaker identification tasks, with performance superior to traditional methods.

3 Strengths and Limitations of SOTA Models

3.1 Strengths

- **Higher Accuracy**: Deep learning models like x-vectors, ECAPA-TDNN, and ResNet-based approaches provide significantly better speaker discrimination power than classical methods.
- **Robust to Noise**: Recent methods that use data augmentation (e.g., adding noise, pitch-shifting) during training make the models more robust to noisy environments.
- **Generalization**: Models like x-vectors generalize well to unseen speakers, as they learn a compact representation of speaker characteristics.
- **End-to-End Learning**: Modern methods based on neural networks can be trained end-to-end, removing the need for feature engineering and manual intervention.

3.2 Limitations

- **Data Hungry**: Deep learning models require a large amount of labeled data to train effectively, which can be a limitation in certain applications.
- **Computational Complexity**: These models are computationally expensive and require powerful hardware (e.g., GPUs) for training and inference, making them less practical for real-time systems in some cases.
- **Overfitting**: If not properly regularized, deep learning models may overfit, especially when the amount of training data is limited.
- **Limited Performance** with Small Datasets: Traditional methods like GMM or i-vectors still perform better on smaller datasets or when few speakers are involved.

4 Evaluation Metrics

4.1 Common Metrics

- **Accuracy**: This is the most common metric used in speaker identification tasks, representing the percentage of correctly identified speakers out of all predictions.
- **Confusion Matrix**: This is used to evaluate how well a model distinguishes between speakers. It provides insights into misclassifications.
- **Precision, Recall, F1-Score**: These metrics are often used to evaluate classification models. They are particularly useful when there's a class imbalance.

4.2 Strengths and Limitations of Metrics

- **Accuracy**: While simple, accuracy doesn't provide much insight into how well the model performs across different classes or speakers.
- **Precision, Recall, F1-Score:**: These metrics provide more insights into model performance, especially when class imbalance is a concern (i.e., some speakers have more samples than others).
- Speaker Identification System Using MFCC (Implemented) – Precision Matrix.

Speaker	Precision	Recall	F1-Score	Support
george	0.98	0.99	0.99	100
jackson	0.98	0.98	0.98	100
lucas	0.98	0.99	0.99	100
nicolas	1	1	1	100
theo	0.97	0.98	0.98	100
yweweler	1	0.97	0.98	100
Accuracy			0.985	600
Macro Avg	0.99	0.98	0.99	600

Figure 4.1: Accuracy Matrix

4.3 Application Overview

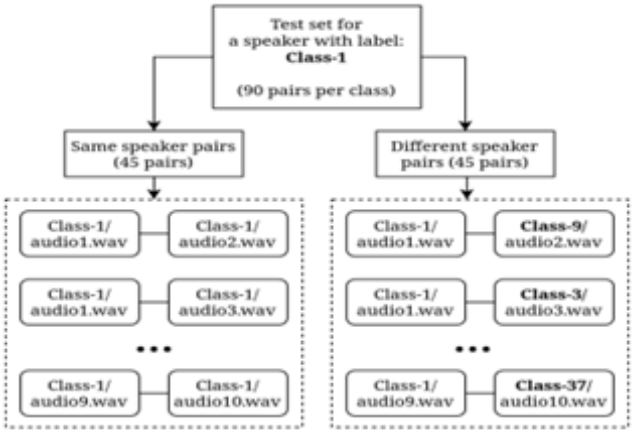

Figure 3. Test set structure for each of the speaker classes in the dataset.

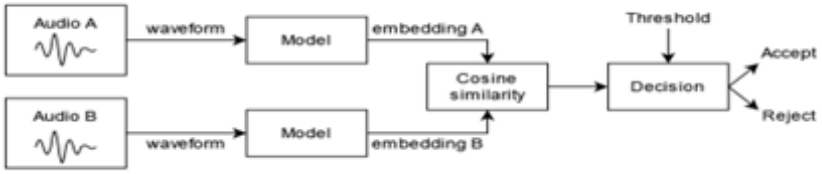

Figure 4. Speaker verification experiment structure.

Figure 4.2: Application Overview

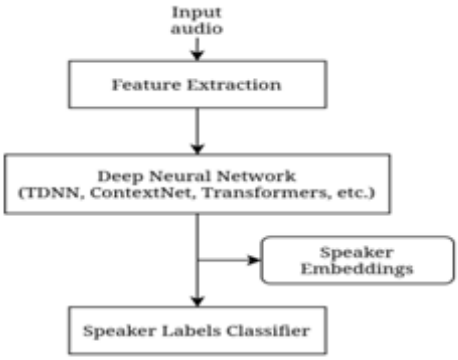

Figure 1. Generalized speaker embedding model architecture.

Table 2. Comparison of the speaker embedding model parameters.

Model	Speaker Embedding Dimension	Training Dataset	Architecture
PyAnnote	512	VoxCeleb	X-vector with SincNet
WavLM	256	LibriSpeech	Transformer
TitaNet	192	VoxCeleb, NIST SRE, Fisher, LibriSpeech	ContextNet with channel attention pooling
Ecapa-TDNN	192	VoxCeleb	Improved TDNN

Figure 4.3: Application Overview

4.4 Fake and Real Prediction Graph

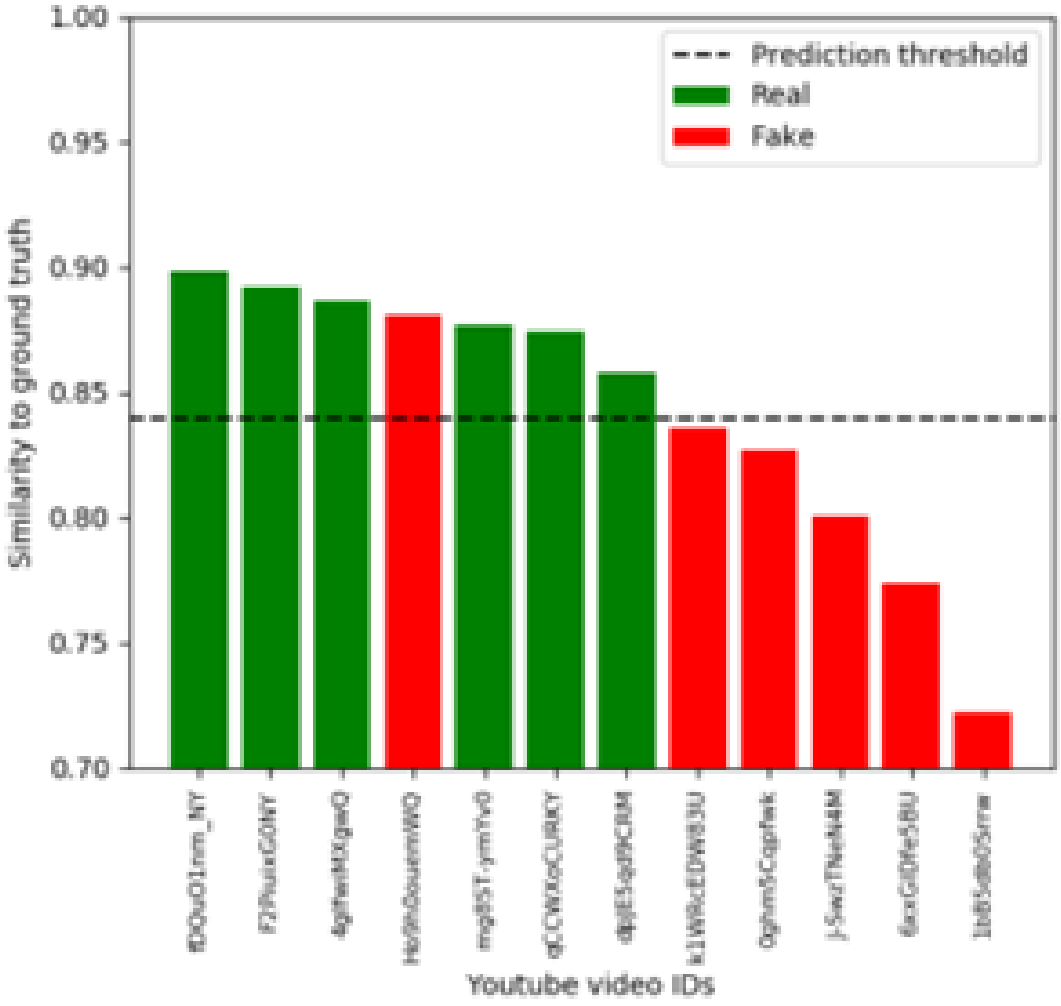

Figure 4.4: Fake and Real Prediction Graph

5 Open Problems and Opportunities

5.1 Open Problems

- **Speaker Variability**: Speaker identification models still struggle with the wide variation in speech characteristics (e.g., accent, age, health status). Modeling such variability is still a challenge.
- **Lack of Labeled Data**: Large-scale, high-quality datasets are expensive and time-consuming to create. Moreover, speaker identification tasks often face data imbalance (certain speakers having more data than others).
- **Real-time Performance**: Real-time speaker identification remains challenging, especially in resource-constrained environments like mobile devices.

- **Generalization Across Domains**: Models trained on one dataset or in a controlled setting may not generalize well to others, especially when environmental conditions (background noise, recording quality) differ.
- **Privacy Concerns**: With the rise of voice biometrics, there are concerns about data privacy, as voice data can reveal sensitive information about individuals.

5.2 Opportunities

- **Cross-Domain Speaker Identification**: Research is needed to develop models that generalize well across different domains, speakers, and environments.
- **Low-Resource Learning**: Techniques like transfer learning and few-shot learning can help build effective models even when labelled data is scarce.
- **Privacy-Preserving Methods**: Exploring techniques such as differential privacy or federated learning could provide solutions for privacy concerns in speaker identification.
- **Hybrid Models**: Combining models from different domains (e.g., combining traditional methods like i-vectors with deep learning models) could improve performance, especially in low-resource settings.
- **Multilingual and Multi-Accent Systems**: Research in multilingual speaker identification can lead to systems that can work across different languages and accents, expanding the usability of these models.

6 Conclusion

Speaker identification has broad applications in security, forensics, and human-computer interaction, with significant advancements in recent years driven by deep learning. State-of-the-art models like x-vectors, ECAPA-TDNN, and transformer-based approaches have outperformed classical methods, providing higher accuracy and robustness to noise. However, there are still challenges related to data availability, real-time performance, and generalization across domains. Future research should focus on improving data efficiency, developing privacy-preserving techniques, and tackling issues like speaker variability and low-resource learning.

References

- Speaker Identification System Using MFCC
- Speech Brain
- Deep Learning Based Speaker Identification System
- A Novel CNN Model for Speaker Recognition
- VoxBlink2: A Large-Scale Speaker Recognition Corpus