Mini Project Report

on

SPEAKER RECOGNITION USING I AND X VECTOR MODELS

Submitted by

Sai Tarun Parasa - 20bcs096

Srihari L - 20bcs129

Venkata Sai Nikhil V - 20bcs135

Vishnukumar P - 20bcs138

Under the guidance of

Dr. Sunil Saumya

Asst. Prof., Dept. of Data Science And Artifical Intelligence

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DHARWAD

15/05/2023

Contents

List of Figures

1 Introduction

Speaker recognition is a technology that aims to automatically identify individuals based on their unique voice characteristics. It has a wide range of applications in various fields, including security systems, forensic investigations, and voice-controlled interfaces. In recent years, the utilization of i-vector and x-vector models has gained popularity in speaker recognition due to their effectiveness in capturing speaker-specific information.

An i-vector is a compact representation of speaker characteristics that can be utilized for speaker recognition tasks. It is derived from a statistical model known as a Gaussian Mixture Model (GMM), which is trained on a set of speaker-specific features extracted from speech signals. The i-vector captures the variability in speaker characteristics while remaining resilient to different recording conditions and channel effects.

On the other hand, the x-vector model represents a more recent advancement in speaker recognition. It employs deep neural networks to extract speaker-specific features from speech signals. The network is trained to classify speakers based on their speech signals, and the output of the network is used as the x-vector. X-vectors have demonstrated superior performance compared to i-vectors in numerous tasks, particularly when dealing with a large number of speakers in the database.

In the context of speaker recognition using i and x-vector models, these models are employed to automatically identify speakers from their voice signals. This entails training the models on a dataset consisting of speech signals from known speakers and subsequently utilizing them to recognize the speakers in a separate test set. The i-vector and x-vector models have exhibited promising results in speaker recognition tasks, and their adoption is expected to continue increasing in the future.

1.1 Problem Statement

"To develop and implement an efficient Automatic Speech Recognition (ASR) system using I-vectors and X-vectors for improving speech recognition accuracy and to compare the efficacy of I - vector and x - vector models in performing speaker recognition tasks."

Automatic Speech Recognition (ASR) plays a crucial role in various applications, including voice-controlled interfaces, transcription services, and language translation. Traditional Automatic Speech Recognition (ASR) systems face challenges in dealing with speaker variations, noisy environments, and large vocabulary sizes. The proposed approach leverages the capabilities of I-vectors and X-vectors, which have shown promising results in capturing speaker-specific information and improving Automatic Speech Recognition (ASR) performance.

1.2 Architecture

1.2.1 Mel Frequency Cepstral Coefficients Architecture

MFCC, which stands for Mel Frequency Cepstral Coefficients, is a widely used feature extraction technique for audio signals in applications such as speech recognition and speaker recognition. The MFCC architecture generally involves the following steps: pre-emphasis, frame blocking, windowing, FFT, Mel-frequency wrapping, logarithmic compression, and Discrete Cosine Transform (DCT). Pre-emphasis is applied to amplify the high-frequency components of the signal, and the audio signal is then divided into frames to obtain time-localized spectral information. Windowing is applied to each frame using a window function such as Hamming or Hanning to reduce spectral leakage at the frame boundaries. The power spectrum of the windowed frame is then obtained using FFT, and the Mel scale is applied to the power spectrum to achieve a logarithmic spacing of the adjacent frequency bands. The logarithm of the Mel spectrum is then taken to compress the dynamic range of the signal, and the DCT is applied to obtain a set of MFCCs that capture the spectral envelope of the signal. These MFCCs can be used as features for various tasks, including speech or speaker recognition.

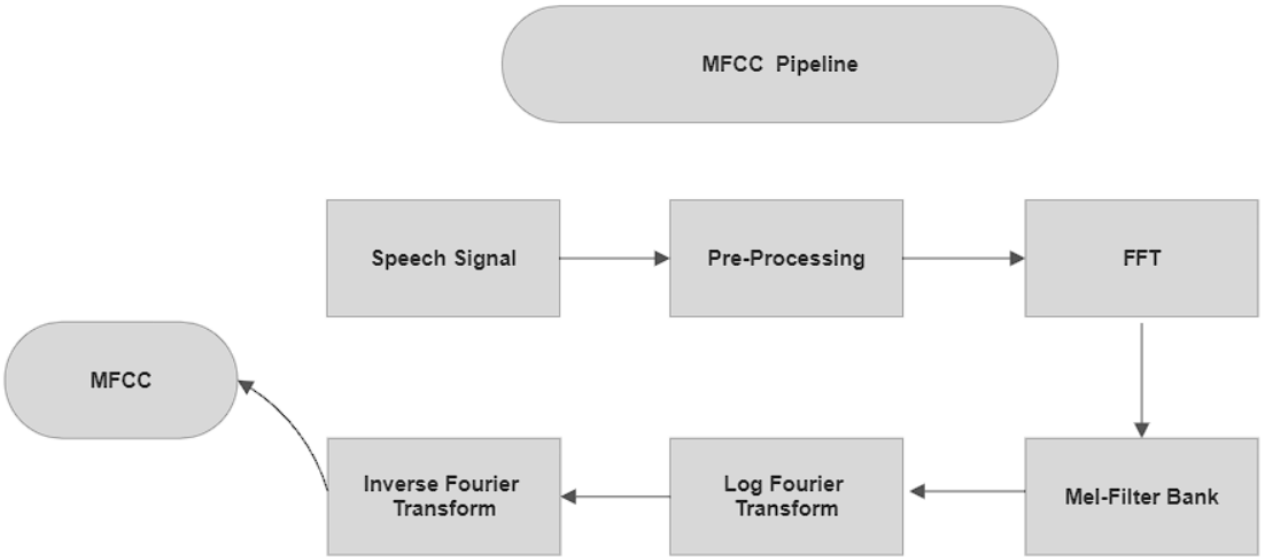

Figure 1. MFCC Block Diagram

1.2.2 I vector and X vector Architecture

The architecture of i-vector and x-vector models in speaker recognition consists of multiple components working together to extract speaker-specific features from speech signals and perform classification tasks.

The i-vector model utilizes a GMM-UBM system to extract low-dimensional representations of speaker-specific information from speech signals. The GMM-UBM system initially trains a GMM on a collection of speech segments from a large number of speakers, creating a universal background model. Subsequently, a speaker-specific GMM is trained by adapting the universal model to each individual speaker's speech. The i-vector is derived from the speaker-specific GMM and represents the speaker-specific information in a reduced-dimensional space. This i-vector is then utilized as input to a classifier, such as a Support Vector Machine (SVM), for speaker recognition.

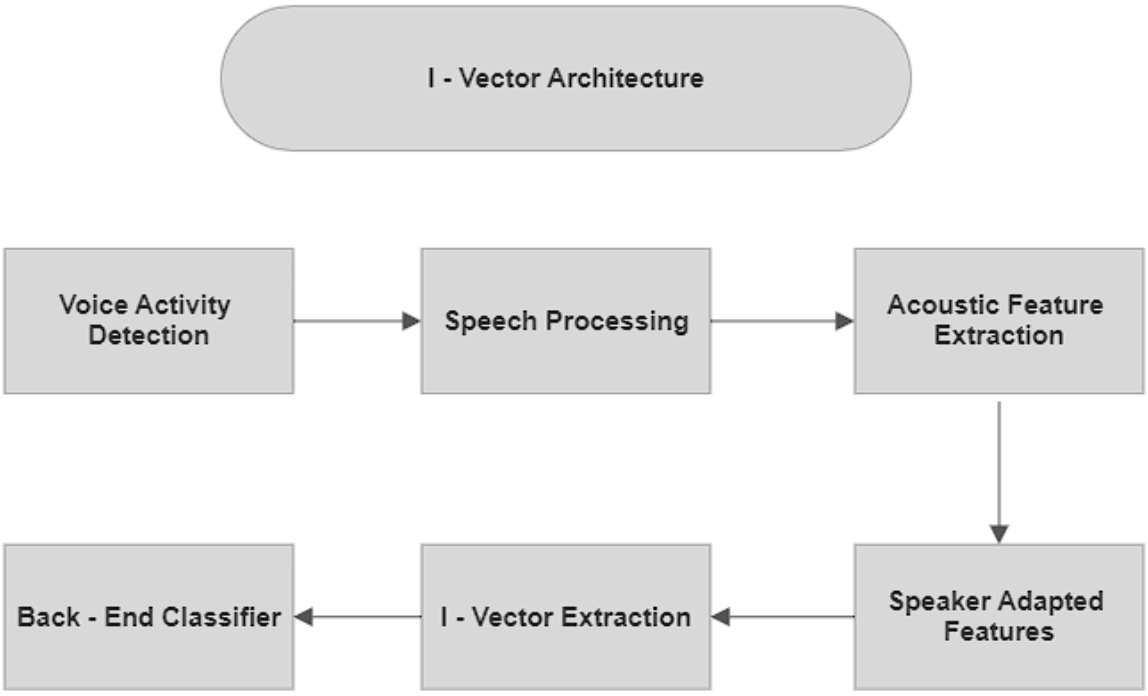

Figure 2. I Vector Block Diagram

On the other hand, the x-vector model employs a deep neural network (DNN) to extract speaker-specific features from speech signals. The DNN is typically trained using a substantial dataset of speech signals and associated speaker labels, enabling it to learn a representation that captures speaker-specific characteristics. The output of the DNN is the x-vector, which represents the speaker-specific information in a high-dimensional space. Similar to the i-vector model, the x-vector is fed into a classifier, such as a neural network or SVM, for speaker recognition.

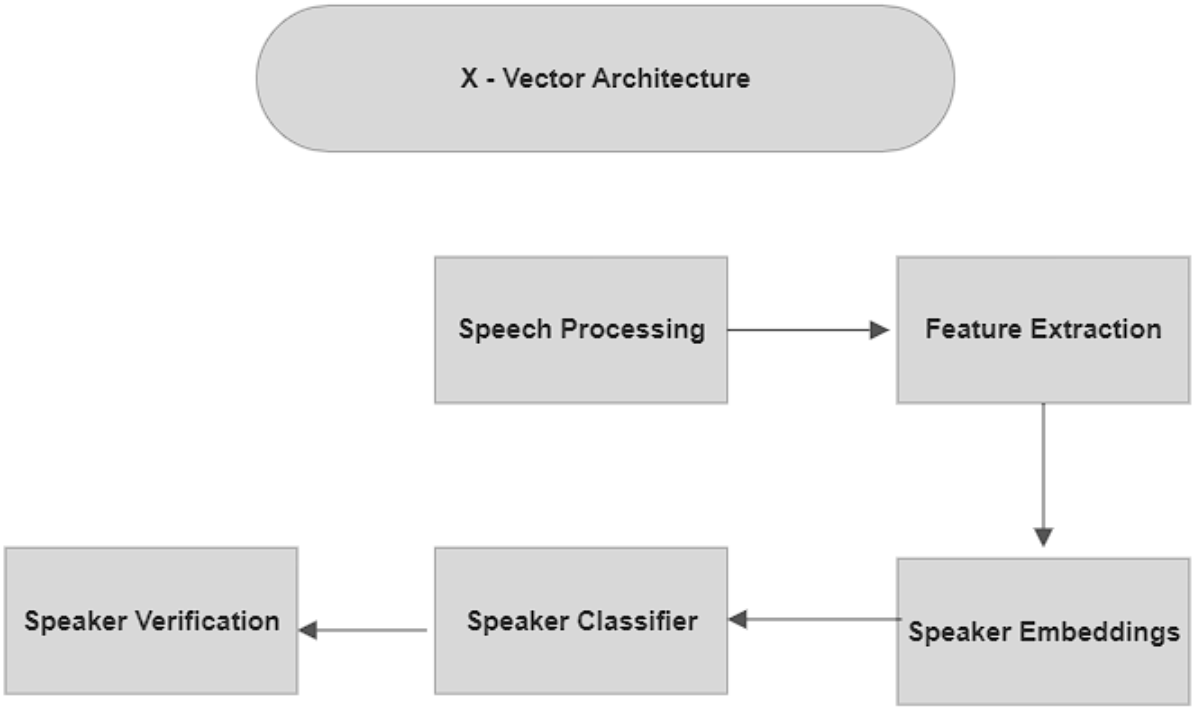

Figure 3. X Vector Block Diagram

Both the i-vector and x-vector models encompass several stages, including feature extraction, model training, and classification. The primary distinction between the two lies in the manner in which speaker-specific information is captured. The i-vector model relies on a statistical model based on GMM-UBM, while the x-vector model utilizes a deep neural network.

2 Related Work

Speaker recognition using I - vector and X - vector models has garnered significant attention in recent years, with numerous studies investigating their effectiveness in various scenarios and comparing their performance against other speaker recognition techniques. In a notable study conducted by David Snyder et al., the focus was on comparing the performance of I - vector and X - vector models in speaker verification tasks using a dataset consisting of speech signals from 50 speakers. The objective was to determine which model exhibits superior performance in terms of metrics such as equal error rate (EER) for accurate speaker identification. The findings of the study indicated that the X - vector model outperformed the I - vector model, showcasing its potential as a more effective approach for speaker verification. Specifically, the X - vector model demonstrated a lower EER, implying that it achieved higher accuracy in distinguishing between genuine and impostor speakers. This outcome suggests that the x-vector model captures and utilizes more discriminative speaker-specific information present in the speech signals.Moreover, the study observed that the X - vector model displayed enhanced robustness in the presence of noisy speech signals. Noisy environments often pose challenges in accurately recognizing and verifying speakers. However, the X - vector model exhibited a greater ability to handle such adverse conditions, making it a more reliable choice in real-world scenarios where speech signals may be corrupted by noise.

In addition to the individual exploration of I - vector and X - vector models, hybrid models that combine the strengths of both have garnered attention in the field of speaker recognition. A study conducted by Sun et al. (2020) proposed a hybrid model that aims to leverage the advantages of both I - vector and X - vector models to achieve even better performance in speaker recognition tasks. The findings of this study revealed that the hybrid model surpassed both the I - vector and X - vector models in terms of EER and accuracy. By integrating the complementary features and capabilities of I - vectors and X - vectors, the hybrid model demonstrated improved accuracy and robustness in speaker recognition tasks.

Overall, the research conducted by David Snyder et al. and the study by Sun et al. shed light on the performance and potential of I - vector, X - vector, and hybrid models in speaker recognition tasks. These findings underscore the significance of exploring and implementing advanced modelling techniques to enhance the accuracy, robustness, and reliability of speaker verification systems in various real-world applications.

3 Methodology

The methodology for speaker recognition using i-vector and x-vector models and comparing their error rates typically involves a series of well-defined steps. Here is an accurate description of each step:

3.1 Preview of Dataset

We had used the Indic Multilingual Speaker Verification (IMSV) Baseline dataset which is a valuable resource commonly used in Automatic Speech Recognition (ASR) studies that involve I - vector and X - vector models. This dataset focuses on speaker verification tasks in a multilingual setting, making it particularly relevant for analysing and improving speech recognition systems in diverse linguistic contexts. When employing I - vector and X - vector models in ASR using the IMSV Baseline dataset, we can aim to develop accurate and robust speaker verification systems capable of identifying speakers across multiple languages. The dataset consists of speech signals from a wide range of speakers, representing different languages, accents, and dialects. To utilize the dataset effectively, we followed a systematic methodology.

1. We prepared the IMSV Baseline dataset by splitting it into training and testing sets, like (dev, enrol, test sets) ensuring a balanced representation of speakers and languages in each subset. This partitioning enables model training and evaluation on distinct data samples.

2. Feature extraction techniques, such as Mel Frequency Cepstral Coefficients (MFCCs), are applied to the speech signals. These features capture relevant acoustic information, representing the short-term power spectrum of sound and providing a basis for subsequent analysis.

3. The I - vector model, a Gaussian Mixture Model-Universal Background Model (GMM-UBM) system is trained on the IMSV Baseline training set. This initial model captures the universal characteristics of the speech data. Subsequently, a speaker-specific GMM is created by adapting the universal model to each individual speaker, generating I - vectors that encode speaker-specific information in a low-dimensional space.

4. The X - vector model employs a deep neural network (DNN) to extract speaker embeddings from the MFCC features. The DNN is trained on the IMSV Baseline training set, using speaker labels to classify and differentiate speakers. The output of the DNN is the x-vector, a high-dimensional representation capturing distinctive speaker characteristics.

5. Both the I - vector and X - vector models undergo testing using the IMSV Baseline testing set. The models are evaluated based on metrics such as equal error rate (EER), which quantifies the accuracy of speaker verification by measuring the trade-off between false acceptance and false rejection rates. Additionally, other performance metrics, such as detection error trade-off (DET) curves, may be employed to assess the models' robustness and discriminatory power.

The use of I - vector and X - vector models with the IMSV Baseline dataset allows us to compare their performance, identifying which model yields superior results in multilingual speaker verification tasks. These evaluations provide valuable insights into the strengths and weaknesses of each model and facilitate further advancements in ASR systems for diverse linguistic contexts. By leveraging the IMSV Baseline dataset and employing I - vector and X - vector models, we can aim to enhance the accuracy, robustness, and generalizability of speaker verification systems in multilingual speech recognition applications.

3.2 Acoustic Feature Extraction

Features are extracted from the speech signals to capture relevant information for speaker recognition. Mel-Frequency Cepstral Coefficients (MFCCs) are features which are used for both models. These coefficients are obtained by applying a Fourier transform to short-time speech segments and extracting the cepstral coefficients that represent the spectral envelope.

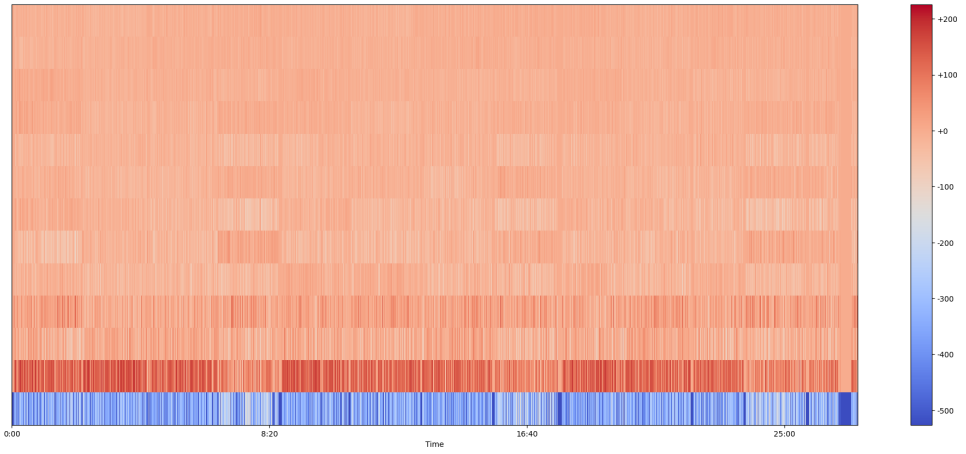

Figure 4. Mel Frequency Cepstral Coefficients Spectrogram

Mel-frequency cepstral coefficients (MFCCs) are a compact and discriminative representation of the speech signal that captures both the spectral and temporal characteristics of the signal. The MFCCs are computed by taking the Fourier transform of the speech signal and mapping it onto the Mel scale, which is a non-linear frequency scale that is more closely aligned with the way humans perceive sound. The log magnitude of the Mel-scaled signal is then passed through a filter bank that mimics the human auditory system, and the resulting filter bank energies are decorrelated using a discrete cosine transform (DCT) to obtain the final MFCC vector representation.

The Obtained MFCC vector representation of 1 Speaker from IMSV Baseline dataset looks like:

```
[[-5.2634235e+02 -5.2634235e+02 -5.2634235e+02 ... -3.9722064e+02
  -4.2360873e+02 -4.4142349e+02]
 [ 0.0000000e+00  0.0000000e+00  0.0000000e+00 ...  1.1566376e+02
   8.3005783e+01  6.2040443e+01]
 [ 0.0000000e+00  0.0000000e+00  0.0000000e+00 ... -1.3848310e+00
  -2.9651266e+01 -4.3139503e+01]
 ...
 [ 0.0000000e+00  0.0000000e+00  0.0000000e+00 ... -1.2258337e+01
  -5.2288551e+00  2.3576517e-01]
 [ 0.0000000e+00  0.0000000e+00  0.0000000e+00 ... -1.7672588e+01
  -1.6527439e+01 -1.1315899e+01]
 [ 0.0000000e+00  0.0000000e+00  0.0000000e+00 ... -1.8253284e+01
  -1.0442340e+01 -4.9509106e+00]]
```

Figure 5. Vector Representation of MFCC's

The MFCC vector captures the most relevant information about the speech signal in a compact and efficient manner, making it a popular choice for feature representation in ASR systems.

3.3 Extracting I-vectors

The I - vector Extraction using IMSV Baseline dataset involves several steps. The basic idea behind the I -vector model is to extract a fixed-length vector representation of a speaker's speech utterance that captures the speaker's characteristics in a compact manner.

1. Voice Activity Detection (VAD): This step involves segmenting the speech signal into non-speech and speech regions. VAD is used to detect the speech regions in the signal.

2. Speech Pre-processing: In this step, the speech signal is pre-processed to remove noise and compensate for the effects of the transmission channel. This includes steps like removing silence regions, removing low-frequency noise, and applying spectral equalization.

3. Acoustic Features Extraction: In this step, features are extracted from the speech signal. These features are used to represent the speech signal in a compact and discriminative way. Typically, Mel-frequency cepstral coefficients (MFCCs) or filter bank energies are used as features.

4. Speaker-Adapted Features: In this step, speaker-specific information is captured by transforming the acoustic features using a speaker-specific model. This is achieved by using a Gaussian Mixture Model-Universal Background Model (GMM-UBM) system. The GMM-UBM is trained using a large dataset of speech from multiple speakers. The GMM-UBM is used to estimate the distribution of speech features for each speaker.

5. I-Vector Extraction: The speaker-specific information captured in the speaker-adapted features is further compressed into a fixed-length I - vector. The I - vector is computed using factor analysis on the speaker-adapted features. The I - vector captures the speaker's unique characteristics in a compact manner.

6. Using Back-End Classifier: Finally, the I - vector is used as input to a back-end classifier. The back-end classifier is trained to classify the speaker as genuine or impostor. This is achieved by using a Support Vector Machine (SVM) or a Probabilistic Linear Discriminant Analysis (PLDA) classifier.

3.4 Extracting X-vectors

X-vectors are high-dimensional vector representations that encode speaker characteristics in a compact and discriminative manner. The X-vector Extraction steps for computing the IMSV Baseline dataset are as follows:

1. Voice Activity Detection (VAD): Similar to I-vector extraction, the first step involves segmenting the speech signal into non-speech and speech regions using Voice Activity Detection.

2. Speech Pre-processing: The speech signal is pre-processed to remove noise and compensate for the channel effects. This includes steps like silence removal, noise reduction, and channel normalization to improve the quality of the speech signal.

3. Acoustic Features Extraction: Features such as Mel-frequency cepstral coefficients (MFCCs), filter bank energies, or other spectral features are extracted from the pre-processed speech signal. These features capture the spectral characteristics of the speech signal and are used as input for the X-vector model.

4. Deep Neural Network (DNN) Training: A deep neural network is trained using the extracted acoustic features. The DNN is typically a time-delay neural network (TDNN) or a convolutional neural network (CNN). The network is trained using a large dataset of speech signals and speaker labels to learn speaker-specific representations.

5. X-Vector Extraction: Once the DNN is trained, the output of a particular layer in the network is used as the X-vector representation. This layer is typically a bottleneck layer or an embedding layer that captures the speaker-specific information. The X-vector is a high-dimensional representation that encodes the speaker characteristics.

6. X-Vector Normalization: The extracted X-vectors are normalized to remove the session and channel variations. This step ensures that the X-vectors are comparable across different recording conditions and environments.

7. Back-End Classifier: The normalized X-vectors are then used as input to a back-end classifier for speaker verification or identification. Common classifiers used include Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), or Probabilistic Linear Discriminant Analysis (PLDA). The classifier is trained on a labelled dataset to distinguish between genuine and impostor speakers.

4 Software Used

Several software tools can be used for speaker recognition using i-vector and x-vector models. Some of the commonly used software tools are:

1.Kaldi:Kaldi is an open-source toolkit widely used for automatic speech recognition (ASR) research and development. It provides a comprehensive set of tools and libraries that facilitate building state-of-the-art ASR systems. Kaldi is known for its flexibility, efficiency, and extensive support for various ASR techniques.Kaldi incorporates advanced algorithms and methodologies, including hidden Markov models (HMMs), Gaussian mixture models (GMMs), deep neural networks (DNNs), and recurrent neural networks (RNNs). It supports both conventional and neural network-based acoustic modeling approaches.

2.MATLAB: MATLAB is a proprietary numerical computing software that can be used for speaker recognition using i-vector and x-vector models. MATLAB provides a wide range of signal processing functions and machine learning algorithms that can be used for speaker recognition.

3.Python: Python is a popular open-source programming language that can be used for speaker recognition using i-vector and x-vector models. Several libraries such as NumPy, SciPy, and scikit-learn provide signal processing and machine learning functions that can be used for speaker recognition.

5 Results and Discussions

The evaluation of speaker recognition using i-vector and x-vector models involves an analysis of their performance in terms of error rates and identification of the factors that affect the performance of these models.

The performance of these models can be assessed using metrics such as equal error rate and confusion matrices. EER measures the point where the false acceptance rate is equal to the false rejection rate over the threshold values, while confusion matrices show the number of correctly and incorrectly identified speakers.

The EER (Equal Error Rate) is typically reported as a percentage that ranges from 0 percent to 100 percent. A lower EER indicates better performance of the system. However, the acceptable range of EER depends on the specific application and the level of security required.

In some applications, such as access control to a secure facility, a very low EER may be required (e.g., below 1 percent). In other applications, such as forensic analysis of speech recordings, a higher EER may be acceptable (e.g.,up-to 10 percent).

Several studies have shown that the X-Vector model performs better than the I-Vector model, based on our project results of equal error rate values i.e., 11.38 for x-vector and 14.73 for i-vector, hence we can show that the performance of X vector compared to I vector Model is robust in Speaker Recognition Systems and even it is better robust in recording conditions, speaker demographics, and noise levels.

However, the performance of these models can be influenced by several factors such as dataset size, recording conditions, choice of feature extraction method, and training parameters. Therefore, it is crucial to carefully design and execute experiments to ensure a fair and accurate comparison of these models. Further research is necessary to improve the performance and robustness of these models.

6 Conclusion

Speaker recognition using i-vector and x-vector models has become a widely studied topic in speech processing due to its practical applications in fields such as security, forensics, and human-computer interaction. While both models with different strengths, the choice depends on the specific requirements. This study provides an overview of the architecture, literature review, methodology, and i-vector extraction using Kaldi for speaker recognition with i-vector and x-vector models. The methodology involves various steps such as dataset preparation, feature extraction, model training, testing, error rate comparison, and analysis. The performance of the models depends on several factors such as dataset size, recording conditions, and feature extraction method, which must be taken into account during the methodology design and execution to ensure fair and accurate comparisons of error rates. In conclusion, speaker recognition with I-Vector and X-vector models shows great promise with practical applications, and further research is needed to enhance the models' performance and robustness.

7 References

1. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5329-5333, doi: 10.1109/ICASSP.2018.8461375.

2. Parwinder Pal Singh, Pushpa Rani," An Approach to Extract Feature using MFCC ",IOSR Journal of Engineering (IOSRJEN),issued on 2014,PP 21-25.

3. A novel voice verification system using adaptive SVM and GMM-UBM" by J. Kim and Y. Kim. This paper proposes a voice verification system that combines GMM-UBM with an adaptive support vector machine (SVM) classifier and evaluates its performance on the UBM-IMSV dataset.

4. A comparative study of speaker verification systems on the IMS corpus" by J. Villalba, I. Lopez-Moreno, and J. Gonzalez-Rodriguez. This paper presents a comparison of different speaker verification systems on the IMS corpus, including those based on GMM-UBM.

5. H. Taherian, Z. -Q. Wang, J. Chang and D. Wang, "Robust Speaker Recognition Based on Single-Channel and Multi-Channel Speech Enhancement," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1293-1302, 2020, doi: 10.1109/TASLP.2020.2986896.

6. D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey and S. Khudanpur, "Speaker Recognition for Multi-speaker Conversations Using X-vectors," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 5796-5800, doi: 10.1109/ICASSP.2019.8683760.

7. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5329-5333, doi: 10.1109/ICASSP.2018.8461375.