Development of Speaker Recognition Model for Forensic Application

A thesis

Submitted in partial fulfilment of the requirement for the award of the degree of

Doctor of Philosophy

in Engineering

by

Gaurav

Reg. No. 951604006
(Old Reg. No. 901604011)

Under the Supervision of

Dr. Saurabh Bhardwaj
(Professor)

Dr. Ravinder Agarwal
(Professor)

Electrical and Instrumentation Engineering Department

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY, PATIALA

(Declared as Deemed-to-be-University u/s 3 of the U.G.C. Act., 1956)

Punjab (India)-147004

February 2024

Keywords

Signal Processing, Speech Processing, Spectrograms, Speaker Recognition, Speaker Identification, Speaker Verification, Speaker Diarization, Feature Extration, Mel Frequency Cepstral Coefficients(MFCCs), Mel Frequency Differential Power Cepstral Coefficients (MFDPCC), Gamma tone Frequency Cepstral Coefficients (GFCC), Power Normalized Cepstral Coefficients (PNCC), Spectral entropy, Classification, Clustering, Convolutional Neural Network(CNN), Hosted cuckoo optimization (HCO) algorithm, Mask R-CNN classifier, Ant Lion Optimizer(ALO), and Arithmetic optimization algorithm.

Abstract

Voice is a natural communication tool humans use to convey meanings, ideas, opinions, *etc.* In particular, "voice" pertains to any sound generated through the vibration of vocal folds when air pressure is from the lungs. It encompasses various characteristics of the speaker, such as ethnicity, age, gender, and emotions. The utilisation of biometrics, particularly voice recognition, has gained popularity in the realm of security. Beyond facial recognition, distinct features like the retina, iris, and voice can be employed to distinguish individuals. Biometrics can be broadly classified as either physiological or behavioural. Physiological biometrics involve features like the face, finger-print, and iris, while behavioural biometrics encompass voice, keystroke, and signature. Among these, voice recognition is one of the most valuable technologies due to its user-friendly nature, widespread acceptance, and cost-effectiveness. Speaker recognition research has been ongoing for several decades, experiencing significant advancements in signal processing, algorithms, architecture, and hardware. Specifically, voice refers to any sound produced by vocal fold vibration when air from the lungs is under pressure. It carries various traits of the speaker, including ethnicity, age, gender, and emotions. The use of biometrics, including voice recognition, has gained popularity in the field of security. In addition to facial recognition, other unique features such as the retina, iris, and voice can also be used to distinguish individuals. Biometrics can be categorised as physiological and behavioural. Physiological biometrics include features like the face, fingerprint, and iris, while behavioral biometrics include voice, keystroke, and signature. Voice recognition is considered one of the most useful technologies. It is easy to use and implement, widely accepted by users, and cost-effective. Research in speaker recognition has been conducted for several decades and has significantly evolved with advancements in signal processing, algorithms, architecture, and hardware.

Normally, speech samples received for forensic examination and comparison originate from uncontrolled environments. Consequently, models were developed for identification and verification in forensic scenarios. The existing methods do not provide sufficient accuracy and robustness of the speech signal. An efficient Speaker Identification framework based on Mask region-based convolutional neural network (Mask R-CNN) classifier parameter optimised using Hosted Cuckoo Optimization (HCO) is developed to overcome the issues. The objective of the method is "to increase the accuracy and to improve the robustness of the signal".

The use of robust feature extraction significantly enhances the efficacy of forensic speaker verification. Although the voice signal is a continuous one-dimensional time series, most

contemporary models mostly use recurrent neural network (RNN) or convolutional neural network (CNN) models. These models lack the ability to comprehensively depict human speech, rendering them susceptible to speech forgery. Therefore, it is necessary to establish a reliable technique to reproduce the human voice accurately and ensure the genuineness of the original speaker. The proposed method presents a Two-Tier Feature Extraction with a Metaheuristics-Based Automated Forensic Speaker Verification (TTFEM-AFSV) model, which aims to overcome the limitations of the previous models. The TTFEM-AFSV model focuses on verifying speakers in forensic applications by exploiting the average median filtering (AMF) technique to discard the noise in speech signals. Both models' performance validation was tested in a series of experiments. A comparative study revealed the significantly improved performance models over recent approaches.

Speaker diarization is a method of splitting individual speakers in the audio stream so that all the speaker's speeches can be separated in the automatic speech recognition (ASR) transcript. Its unique audio features divide the speakers, and its speeches can be bucketed together. As mass gatherings and communication increase, the process of speaker diarization might add complexity to efforts to enhance the clarity of speech transcripts. In response to these concerns, an automated speaker diarization system has been devised by employing an arithmetic optimization algorithm alongside a deep belief network technique known as ASDS-AOADBN. To address these issues, an automated speaker diarization system using an arithmetic optimisation algorithm with a deep belief network (ASDS-AOADBN) technique is developed. The model's primary purpose lies in identifying and classifying speaker signals from input audio signals. The experimental result analysis stated the better performance of the ASDS-AOADBN technique over recent state-of-the-art DL models.

Table of Contents

List of Tables

List of Figures

Acronyms and Abbreviations

ASR Automatic Speaker Recognition
ALO Ant Lion Optimizer
ANN Artificial Neural Networks
AMF Average Median Filtering
AOA Arithmetic Optimization Algorithm
CNN Convolutional Neural Network
CNN Convolutional Neural Network
DCF Detection Cost Function
DCT Discrete Cosine Transform
DET Detection Error Trade-off
DFT Discrete Fourier Transform
DL Deep Learning
DNN Deep Neural Network
DTW Dynamic Time Warping
EER Equal Error Rate
EM Expectation Maximization
GFCC Gamma tone Frequency Cepstral Coefficients
HCO Hosted cuckoo optimization
IFT Inverse Fourier Transform
LFCC Linear Frequency Cepstral Coefficients
LP Linear Prediction
LPC Linear Prediction Coefficients
LPCC Linear Prediction Cepstal Coefficients
LSTM Long Short-Term Memory
MAP Maximum a Posteriori
MFCC Mel Frequency Cepstral Coefficients
MFDPCC Mel Frequency Differential Power Cepstral Coefficients
MinDCF Minimum DCF
ML Machine Learning
ML Maximum Likelihood
pdf Probability Desity Function
PLP Perceptual Linear Prediction
PNCC Power Normalized Cepstral Coefficients
ReLU Rectified Linear Unit
RNN Recurrent Neural Networks
SNR Signal-to-Noise Ratio
STFT Short Time Fourier Transform
SI Speaker Identification
SD Speaker Diarization
SV Speaker Verification

Declaration

I declare that the thesis entitled **Development of Speaker Recognition Model for Forensic Application** submitted by me for the degree of Doctor of Philosophy, is the record of research work carried out by me during the period from **Jan 2017 to Jan 2024** under the supervision of **Prof. Saurabh Bhardwaj and Prof. Ravinder Agarwal** this has not formed the basis for the award of any degree, diploma, associateship, fellowship, titles in this or any other University or other institution of higher learning.

I further declare that the material obtained from other sources has been duly acknowledged in the thesis. I shall be solely responsible for any plagiarism or other irregularities if noticed in the thesis.

Gaurav
Registration no.: 951604006

Date: 08/02/2024

Certificate

This is to certify that the thesis "**Development of Speaker Recognition Model for Forensic Application**", which is being submitted by **Gaurav** to the Department of Electrical & Instrumentation Engineering, Thapar Institute of Engineering & Technology (Deemed to be University), Patiala, for the award of the Degree of Doctor of Philosophy, is a record of bonafide research work carried out by him under our guidance and supervision and has full-filled the requirements for the submission of this thesis, which to our knowledge has reached the requisite standard.

To the best of our knowledge: (i) the candidate has not submitted the same research work to any other institution for any degree/diploma, Associateship, Fellowship or other similar titles (ii) the thesis submitted is a record of original research work done by the Research Scholar during the period of study under our supervision, and (iii) the thesis represents independent research work on the part of the Research Scholar.

(Dr. Saurabh Bhardwaj)
Professor
E.I.E.D.
T.I.E.T., Patiala

(Dr. Ravinder Agarwal)
Professor
E.I.E.D.
T.I.E.T., Patiala

Acknowledgements

I want to express my profound sense of gratitude and indebtedness to the almighty God for giving me such an opportunity. My PhD journey at TIET would not be more pleasant without the support of my supervisors, colleagues, friends, and family. There are many names to be mentioned who came across during this complex journey. This small piece of document is not enough to express my gratitude to them for their continuous support along the way.

First and foremost, I would like to express my respect and a deep sense of gratitude to my supervisors, **Dr. Saurabh Bhardwaj**, Professor, Electrical & Instrumentation Engineering Department, Thapar Institute of Engineering Technology, Patiala, and **Dr. Ravinder Agarwal**, Professor, Electrical & Instrumentation Engineering Department, Thapar Institute of Engineering Technology, Patiala, for their wisdom, vision, expertise, guidance, enthusiastic involvement, and persistent encouragement during the planning and development of this research work. I also gratefully acknowledge their painstaking efforts in thoroughly going through and improving the manuscripts, without which this work could not have been completed. I would especially like to mention my principal supervisor, Dr. Saurabh Bhardwaj, for his consistent help and support. He has been a perfect mentor for me. I still marvel at his exceptional generosity and encouragement shown towards me. I am grateful to him for giving me directions, confidence and valuable advice and for everything that I have learned from him. He has had a tremendous impact on the way I perceive research and life in general.

I am highly obliged to **Prof. Padmakumar Nair**, Director, Thapar Institute of Engineering Technology, Patiala, for support and necessary facilities to carry out and complete this work on a steady course.

I am also thankful to present Chairman of the Doctoral Committee**, Dr. Sunil Kumar Singla,** Professor & Head, Department of Electrical and Instrumentation Engineering, for the much-needed support throughout the work and for providing all the facilities, help and encouragement for carrying out the research work.

I want to extend my special thanks to **Prof. R.S. Kaler** for his guidance and support.

I am very thankful to **Dr. Mandeep Singh**, Professor, Department of Electrical & Instrumentation Engineering, **Dr. M. D. Singh**, Professor, Department of Electrical & Instrumentation Engineering and **Dr. Vinay Kumar**, Professor, Department of Electronics & Communication Engineering, for being the members of the Doctoral Committee and spending their valuable time reviewing and critically examining the work.

I want to express my loving gratitude to my family members for their patience, love, support, and encouragement during this journey.

I wish to express my appreciation to my friends Sachin Sharma, Gaurav Kumar, Praveen Bhola, Krishan Kumar, Jitender Virk, Anshul Mahajan, Manish Singla, Maan Singh, Rohit Arora, Rajat Singla and Shivam Agarwal and thanks to research fellows at the department for their help and motivation throughout my research work.

I would also like to extend my special thanks to Mrs Poonam, Mr Purushottam, Mr Jaideep, Mr Ganesh, Mr Arvind, Mr Sunil and other staff members of the EIE Department for their timely help and cooperation, extended throughout the course of the investigation.

Of course, a very special thanks to my wife Kritika and my beloved family for their unconditional support and patience, for being by my side in the most stressful and hardest moments, and for enjoying the good ones.

As it is a prolonged journey, maybe I have forgotten to mention a few names; nonetheless, they remain a core part of this mammoth task, and I seek forgiveness for the same and offer my kind respect to every one of them.

Gaurav

1. Chapter - Introduction

Let's take a moment to retrospect upon the events of the day thus far, commencing from the instant of awakening this morning. Depending on the geographical location, it is plausible that one employed a password to deactivate the domestic security system. Perhaps a physical key was utilized to unlock the front door and initiate the vehicular conveyance. In the realm of electronic transactions, the requirement to furnish a security PIN code during the act of purchasing goods or withdrawing funds from an Automatic Teller Machine (ATM) might have presented itself. Alternatively, gaining entry into the professional environment could have involved swiping an access card or undergoing biometric authentication, such as fingerprint scanning. These instances exemplify various manifestations of access controls, an omnipresent facet in contemporary daily existence. A predominant methodology in the domain of access control is the amalgamation of identification and authentication processes. One can use their voice (speech signal) as a single key to access and authorize entry to secured systems or premises. This vocal biometric authentication, utilizing the unique characteristics of one's speech signal, represents an evolving frontier in access control methodologies. By employing advanced algorithms to analyze and verify the distinctive vocal patterns of individuals, this technology enhances the precision and security of authentication processes. The utilization of voice as a singular key not only streamlines access but also introduces a layer of convenience in interactions with security systems. As technology continues to advance, the amalgamation of identification and authentication processes, with innovations such as vocal biometrics, contributes to the ongoing refinement of access controls, ensuring a harmonious balance between security and user-friendly experiences in dynamic contemporary landscape.

The term "speech signal" pertains to the auditory representation of spoken communication. The articulatory organs move in a certain way to create sound when we speak., including the vocal cords, tongue, and lips, and is then transferred as sound waves through the air. These sound waves possess distinct characteristics, such as amplitude, frequency, and duration. Throughout history, speech has evolved as the primary mode of human communication. In recent decades, experts have dedicated their efforts to researching the human voice and speech due to its fundamental and captivating nature.

The voice signal is an intricate waveform that may be examined and quantified using many methodologies. Two methods exist for studying speech signals: one uses time-domain analysis to capture the signal's temporal qualities, while the other uses frequency-domain analysis to

analyse the signal's frequency content. The characteristics of a speech signal include its pitch, intensity, duration, and formants. These qualities are essential for transmitting linguistic information and enabling listeners to hear and comprehend speech. Furthermore, the voice signal may exhibit variations depending on characteristics such as speaker identification, emotional state, accent, and background noise. The fluctuations in the speech signal have a role in the distinctiveness of individual voices and may be used for the goal of identifying speakers.

1.1 Motivation and Overview

People use a wide variety of communication strategies in everyday life, including verbal expression, nonverbal cues, visual cues, and body language. Nonetheless, voice remains the most common means of human communication among these kinds. Because of its unique personality and depth of meaning, spoken word has long been considered the most powerful art form. Rich aspects include not just the spoken text (words), but also the speaker's gender, emotions, health, attitude, and identity, as shown in Figure 1.1. For effectual communication, such information is very significant. Humans are unique in their ability to transmit information with their voices. Opinions and sentiments can be effortlessly expressed through speech communication. As a result of developments in human-computer interaction systems, it is now possible to digitise acoustic speech signal information for use in biometric security systems and even the management of enormous archives of spoken audio databases. The vast amounts of biometric data sent by acoustic speech signals may be combined with a wide range of biometric security systems. These systems can scan retinas, fingerprints, faces, veins, *etc.*

Speaker Recognition is the automated identification of individuals based on unique characteristics present in voice signals. The speaker's voice can be used to verify their identity and manage their authorization for services like voice dialling, voice shopping, voice banking, information services, voice messaging, database access, security control for restricted areas, and remote computer access when this technique is applied.

A speaker recognition system's first step is to try to imitate an individual's vocal tract features [1]. This can mean a statistical model that mimics the human vocal tract in terms of output or a mathematical model that depicts the physiological system responsible for the human voice [2]. Once a model is constructed and linked to a person, additional instances of speech may be evaluated to determine the probability of their being produced by the specific model in question, as opposed to other observable models. This serves as the fundamental approach for all speech detection applications. Speaker recognition is significant because it is the only

biometric that can be remotely tested (identified or validated) using the current telephone network infrastructure, just relying on the speaker's identity and voice signal. Speaker identification is very important and unmatched in several practical applications [3]. Speech may be described in terms of the message information it carries from a signal-processing perspective. Speech signals may provide three main forms of information: Speech content (text), Language, and Speaker Identity [4]. It is worth noting that as the number of mobile phones continues to increase and their complexity continues to develop, speaker identification will become more important in the near future.

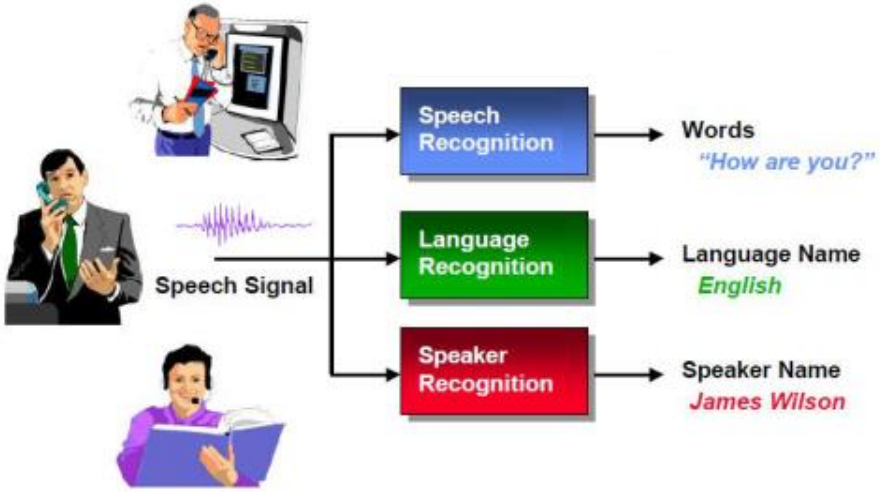

Figure 1.1: Information from speech [4]

In the speaker recognition task, a significant enhancement in performance has been noted since the implementation of deep neural networks (DNNs) in recent times [5]. In contrast, the development of reliable speaker models via acoustic DNN training necessitates a substantial volume of speech data that has been accurately transcribed. The focus of this is on the development of a speaker recognition system for forensic applications. The biometric speaker recognition models provide good results for discriminating the speakers under controlled conditions. In forensic settings, however, the conditions in which questioned speaker recordings are made remain largely uncontrolled.

1.2 Speaker recognition principles

Every speaker possesses distinct qualities when it comes to speaking, such as a unique dialect, manner of speaking, diction, cadence, vocabulary usage, and grammar. Speech serves as an exact indicator for biometric security systems [6]. Identification, detection/verification, and segmentation and clustering are the three distinct tasks that may be separated from the overall field of speaker recognition [7], depending on the application. The following provides a concise

summary of the information obtained from speech signals, specifically in terms of extracted features.

1.2.1 Features

Feature extraction is the initial and foundational stage of the speaker recognition system. It eliminates redundancy and extracts speaker-specific information from the original waveform. A series of feature vectors are generated from the unprocessed speech signal during the feature extraction procedure. These feature vectors highlight distinctive information about the speaker while removing redundant data. Every individual possesses distinctive trademark qualities in their vocal expression. Distinguishing the distinct qualities of individual speakers proves to be a challenging task. These qualities are exceptional due to the fact that distinct vocal tract physiologies and culturally influenced speaking behaviours characterise each speaker. The existence of vocal contrasts among indistinguishable siblings is noted, notwithstanding their comparable vocal tract morphology and acoustic properties. To recognise different voices in speaker reognition, it is necessary to take into account quantifiable and distinct characteristics of speech, regardless of whether the recognition is performed by humans or machines. The aforementioned distinctive attributes are denoted as the features. The optimal features would possess the qualities listed below [8] :

a. Variability between speakers should be considerable, while variability within speakers should be minimal
b. Impervious to mimicry attempts
c. The extraction process for the speech signal should be straightforward
d. It needs to be a regular and organic part of the speech signal.
e. Outperform in environments with high levels of noise
f. The health and mood of the speaker should not influence speaker-dependent information
g. It must to be resilient over a variety of transmission

There are two types of speaker-specific information: "high-level" features and "low-level" features [9]. The vocal fold size and other anatomical characteristics make up the "low-level" traits, which are associated with the vocal tract. Conversely, the speaker's prosody, education, location of birth, accent, and pronunciation are examples of the "high-level" attributes [10]. The amount of time required to extract and process these two types of features differs. The low-level features are computed within a brief time frame, usually ranging from 10 to 30

milliseconds, and possess a very simple structure. The voice signal is potentially vulnerable to a wide variety of external variables in addition to the inherent content diversity that it has. It is thus reasonable to anticipate that the voice signal will display the features listed below.

- quasi-stationarity (10-40ms)
- wide dynamic range (30-80dB)
- voiced, unvoiced and silence intervals
- possible mimicry by imposters
- variability due to the speaker's health

Environmental additions:

- white, coloured or impulsive noise
- possible effects of speech coding
- channel and recording disturbances
- co-talker interference or room reverberation
- band limiting
- temporal channel variability

The task of determining appropriate signal processing algorithms for speaker recognition necessitates a delicate equilibrium, as it entails adjusting the relative importance of precision, efficiency, and robustness in pursuit of the most optimal system. Prior to 1990, the predominant body of research in this field had concentrated on the improvement of precision and productivity. This was owing to the fact that, in comparison to their current selves, computer processing capabilities were comparatively limited. Recent research on Speech Signal Processing has focused on enhancing precision and robustness in tandem with the implementation of novel methodologies. In cases when a significant amount of voice data over an extended duration is necessary to extract advanced features. Below is a concise summary of these features.

1.2.1.1 Short-term spectral features

The non-stationary character of the voice signal necessitates its division into frames of 10-30 ms. It is believed that the signal is quasi-stationary in this frame. Each frame has its spectral properties retrieved. Utilising smooth window operations, each frame is pre-emphasized and multiplied. An efficient method for implementing DFT, the fast Fourier transform (FFT) separates a signal into its frequency components [11] [12] [13]. DFT's magnitude spectrum shape is most relevant in speaker recognition because it reveals vocal tract resonance features.

The magnitude spectrum is processed by a bank of band-pass filters. Typically, the filters have a triangle shape. Fewer filters are used for lower frequency ranges and more for higher frequency ranges. After that, Mel-frequency cepstral coefficients (MFCC) became the cutting edge of many speech processing applications, such as emotion detection, speaker and language recognition, speech recognition, and many more. In addition to MFCC, the following speech aspects pertaining to the spectrum have also been dominant in speaker recognition: A number of linear prediction coefficients, including LPCs, LPCCs, LSFs, and PLP coefficients (perceptual linear prediction) [14] [15] [16] [17].

1.2.1.2 Dynamic features

One factor that helps speaker recognition is the time-varying nature of the data in the voice stream. A common way to incorporate time-related data into features is to use the features' velocities (often called Δ features) and accelerations (also called ΔΔ features) [18]. Dynamic features are computed by combining the base coefficients with the features' velocities and accelerations at the frame level [19]. We find these accelerations and velocities by measuring the time difference between neighbouring feature coefficient vectors. The usage of a 12-dimensional MFCC with Δ and ΔΔ coefficients, for instance, suggests that each frame will include 36-dimensional characteristics.

1.2.1.3 Prosodic features

An inherent limitation of spectral features, such as MFCCs, is their susceptibility to the influence of noise signals. It lacks details on phonemes and just focuses on the amplitude spectrum of the speech signal, disregarding the phase spectrum. Prosodic characteristics include extended units such as syllables, words, and utterances. Prosody encompasses the elements of intonation, rhythm, formant frequencies, pitch, and syllable length, in speech. Speaker-to-speaker variation is seen in prosody [20].

The fundamental frequency ($F0$) is the paramount prosodic characteristic. The integration of spectral characteristics with $F0$ correlated data demonstrates superior performance, particularly in noisy environments [21]. Additional prosodic characteristics include pitch, energy allocation, duration, and speaking pace.

The *F0* represents the size of the larynx, whereas temporal fluctuations are associated with the manner in which speech is delivered. Therefore, $F0$ encompasses both the behavioural and physiological traits of the speaker. The local and longer-term temporal information of $F0$ may be used to model and extract multidimensional pitch and other properties [22].

1.2.1.4 High-level features

Higher-level features include phonetic and lexical information. Additionally, it encompasses speaker attributes derived from pronunciation, prosody, and lexical data. The voice signal contains a significant amount of long-range and linguistic information. The integration of higher-level information with lower-level cepstral information may greatly enhance the performance of speaker recognition systems [23]. High-level features assess several characteristics of an individual's speech, such as the choice of words, accent, and duration of pauses between words. High-level features capture the conversational attributes of speakers, such as their usage of certain words ("ya", “hmm”, "okay", "aww", "got it", *etc.*).

Higher-level information has the potential for enhanced resilience to channel variant since factors like lexical use or temporal patterns remain consistent despite fluctuations in auditory settings [24]. More advanced features may provide valuable information about a speaker, including the subject being addressed, the speaker's interaction with another individual, their emotional state, and any speech disfluencies.

1.2.2 Outline of speaker recognition technology

Speaker recognition is a crucial component of a biometric system, particularly in situations when direct physical contact for detection is not feasible. Speech serves as a potent means of identifying a person's identity in mobile/telephone apps. Furthermore, speech signals are somewhat more accessible compared to other forms of biometrics. The speaker recognition issue involves finding the unknown speaker's identity, and the speaker verification problem where the claimed identity of the speaker is verified. Both problems (identification and verification) are inherent in speaker recognition. The three primary parts of any speaker recognition system are feature extraction, speaker modelling, and score calculation. In literature, the process of automatically identifying and verifying speakers is referred to as speaker recognition. The first phase of extracting relevant features is known as the front end, while the subsequent stages of modelling and calculating scores are referred to as the back end. Figure 1.2 illustrates the process of speaker recognition and highlights its essential characteristics and models. The next subsections provide a comprehensive overview of each module and the methods to enhance its resilience.

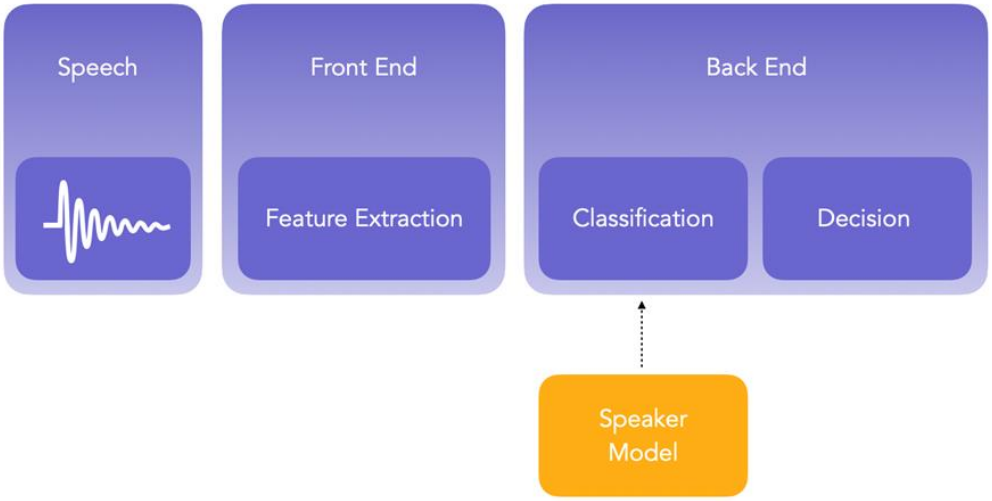

Figure 1.2: Overview of speaker recognition system

1.2.2.1 Speaker-specific speech production system attributes

Speech is created by the airflow generated by a sound source as it travels through the vocal tract filter, according to the principles of the fundamental source-filter model. The ability to identify speakers from speech relies on the influence of human speech organs on speech production. The physical characteristics of speech organs and the variety in behaviour, such as style and accent, provide distinct information that is particular to each speaker during speech [25]. The configuration of the vocal tract is indicative of the speaker's unique qualities in the range of speech frequencies, while the vocal folds specifically provide information about the speaker's age and gender. Vocal tract length has a substantial correlation with lower formant frequencies, which in turn is connected with body size. Hence, the overall configuration of the vocal tract transmits speaker characteristics via the lower frequencies of the formant. The vocal cord and vocal tract, among other speech organs, contribute to the unique characteristics of a speaker's speech. However, it is the speech organs that remain consistent regardless of the exact sounds produced that are crucial for identifying a speaker [25]. Phonetic variety is a challenge. The structure of the hypopharynx or hypopharyngeal cavities remains stable throughout vowel production, but it might vary significantly among different speakers.

1.2.2.2 Front-end processing

It is possible that not all data included in voice signals is necessary to distinguish speakers from the overall population. In order to construct reliable speaker models, it is essential to use just the part of the speech signal that contains information particular to the speaker for the speaker recognition system to function. Speech activity detection and feature extraction are part of the front-end processing that starts speaker identification systems.

1.2.2.3 Voice activity detection

Voice Activity Detection (VAD) is essential for distinguishing between speech and non-speech portions of a voice signal, hence improving the accuracy of the speaker recognition system. The widely used VAD characteristic is energy-based, which leverages the energy of the spectrum, assuming that speech amplifies the signal energy in comparison to background noise or silence. Typically, this distinction is made by using low-pass energy for speech parts and high-pass energy for noise/silence portions [26]. In addition, features such as discrete Fourier transform coefficients, mel-scale filter bank energies, mel-frequency cepstral coefficients, and signal-to-noise ratio (SNR) are also used [27] [28].

One characteristic often employed in vocal activity detection (VAD) is the zero-crossing rate, which counts the number of times a signal goes from positive to negative amplitudes. Additional features include linear predictive speech analysis, entropy, and the variability of phoneme probability distribution as determined by a multilayer perceptron [29]. After extracting the features, it is essential to make VAD judgements in order to differentiate between speech and non-speech parts. One way to do this is by using direct feature-thresholding approaches or modelling features, such as utilising Gaussian Mixture Models (GMM) for likelihood ratio thresholding [30]. Certain systems use pre-trained generative Gaussian models to enhance signal modelling. Statistical categorization entails the use of Hidden Markov Models (HMM) with temporal restrictions to divide speech and non-speech data into segments [31]. These technologies jointly enhance the detection of vocal activity and its uses in speaker identification systems.

1.2.2.4 Feature extraction

Speech feature extraction is the creation of a human model using concise snippets of information. The process entails analysing a speech signal over a duration of 20-30 milliseconds since the vocal tract remains mostly unchanged during this period [32]. Pre-emphasis amplifies high-frequency elements by using a high-pass filter. The Mel frequency cepstral coefficient (MFCC) [14], linear prediction cepstral coefficients (LPCC) [15], and perceptual linear prediction (PLP)-based parameters [16] are important features used to analyse short-term characteristics in speaker recognition [17].

MFCCs extract the magnitude of the spoken signal from an auditory standpoint. Hamming and Hanning windows are used to extract MFCC features. Following the conversion of the short-term signal to the Fourier domain, the amplitude spectrum is subjected to triangle Mel scale

filters. The Mel scale establishes a relationship between a frequency in Hz (f_{Hz}) and a Mel scale frequency f_{Mel}, based on the perception of diverse tone heights by the human ear, as seen in the equation (1.1).

$$f_{Mel} = 2595.log\left(1 + \frac{f_{Hz}}{700}\right) \quad (1.1)$$

Ultimately, the discrete cosine transform is used to remove the correlation between the log energies of the Mel-filtered spectrum, resulting in the extraction of MFCC. MFCCs are calculated by applying an acoustically inspired filter bank, followed by logarithmic scaling, the discrete cosine transform (DCT), and finally, selecting a certain number of prominent coefficients.

$$C_n = \sum_{k=1}^{K}[log\, Y(k)]cos[\frac{\pi n}{K}(k - \frac{1}{2})] \quad (1.2)$$

Where, With n being the cepstral coefficient index, Y (k) is the K-channel filter bank's output.

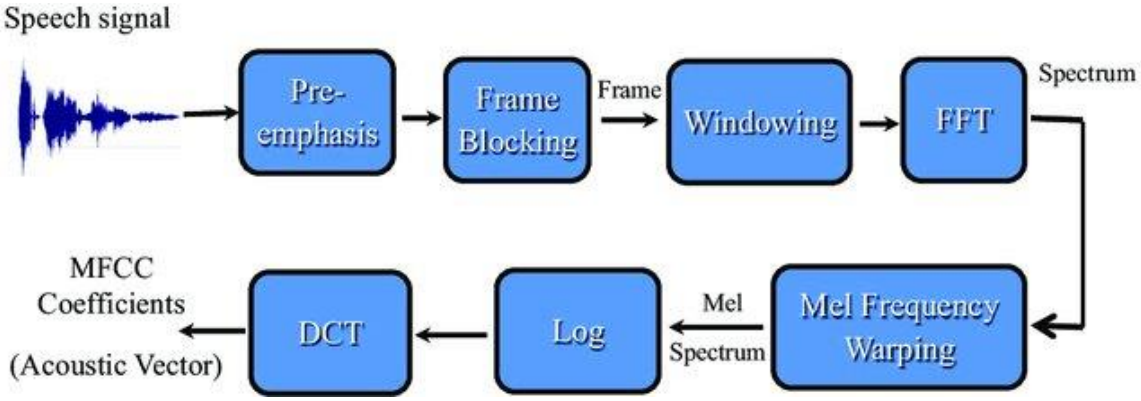

Figure 1.3 MFCC feature extraction [32]

Figure 1.3 depicts the procedure of mining Mel-frequency cepstral coefficients (MFCC). This MFCC feature is used as the baseline feature for the methodologies in this thesis. The vocal tract is represented as a linear filter by using linear prediction analysis to compute LPC coefficients. Speaker identification often uses the coefficient or cepstral domain transform (LPCC) of this filter. The aforementioned characteristics mostly depend on the amplitude of the speech signal, which is only one component of the signal. Phase-based characteristics were found to supplement this information and improve the system's performance in real-world scenarios.

1.2.2.5 Speaker modelling

After excavating the feature vectors from the speech signal, the subsequent step in speaker identification is speaker modelling. Speaker modelling is an essential component of automated speaker recognition systems. Its purpose is to accurately distinguish between speakers by

correctly characterising their sessions and utterances. An effective speaker model streamlines the process of speaker recognition. Nevertheless, in order to account for differences such as channel and noise, it is essential for these modelling systems to possess the ability to differentiate between different speakers efficiently. Gaussian Mixture Models (GMMs) are often employed in speaker verification, particularly for text-dependent tasks [33]. For each speaker model, this modelling technique calculates parameters that are associated with the Gaussian Mixture Model (GMM). The Universal Background Model (UBM) was created as a means to represent the collective speaker population rather than individual speakers, referred to as speaker-independent GMM modelling [34]. Speaker-specific models are often estimated using the Universal Background Model (UBM) via Maximum a posteriori (MAP) adaptation [35].

Kenny *et al.* suggested using joint factor analysis (JFA) as a solution to the issue of considerable speaker data being exposed to the channel domain [36]. Nevertheless, this approach resulted in inadvertent data disclosure. Dehak *et al.* proposed the use of total-variability modelling, which is a method that takes into account the characteristics of both the speaker and the channel [37]. This methodology represents the GMM super-vector in a lower-dimensional i-vector space. After incorporating deep neural networks (DNN) into speaker verification, DNN has taken the position of UBM in order to collect enough posterior statistics for total-variability and i-vector extraction training [38] [39].

1.2.2.6 Convolutional Neural Network (CNN)

One subset of artificial neural networks developed specially tailored for the purpose of processing and analysing visual input is the Convolutional Neural Network (CNN). Its effectiveness in applications like image/visual identification and classification has been shown since it can capture hierarchical characteristics by using convolutional layers. It is believed that convolutional neural networks (CNNs) take their cues from a neuron's receptive field, which consists of the specific areas of a visual field to which the neuron is sensitive. Figure 1.4 represents the general CNN architecture for spectrograms.

The constituents of a Convolutional Neural Network (CNN):

a. *Convolutional layers* use convolutional procedures to extract features from incoming data. Convolution entails the process of systematically moving a diminutive filter, also known as a kernel, across the input data, therefore collecting localised patterns or

features. The outcome is a feature map that accentuates pertinent geographical information. Mathematically, the convolution operation can be represented as:

$$(f * g)(t) = \sum_a f(a).g(t-a) \quad (1.3)$$

where,

- f is the input signal (*e.g.*, a spectrogram representation of speech),
- g is the filter (also known as a kernel),
- $*$ refers to the convolution operation,
- t represents the position in the output feature map.

b. *Pooling layers* reduce the size of the feature maps by down-sampling their spatial dimensions. This aids in reducing the computational burden and guarantees that the network is more resilient to fluctuations in input.

The max pooling operation is given by:

$$MaxPooling\ (x,y) = max_{i=0}^{x-1} max_{j=0}^{y-1} Input(i,j) \quad (1.4)$$

Here,

- x and y are the dimensions of the pooling window,
- Input *(i,j)* represents the input value at position *(i,j).*

c. *Fully connected layers* Using FC layes, the network is able to build abstractions and predictions at a high level by connecting neurons in one layer to neurons in the layer below it. Let x_i be the output of the previous layer, w_{ij} be the weight connecting neuron i to neuron j, and b_j be the bias for neuron j. The output y_j of neuron j in the fully connected layer is given by:

$$y_j = \sigma\ (\sum_i w_{ij}.x_i + b_j \quad (1.5)$$

where,

- σ is the activation function (e.g. ReLU or Sigmoid),
- i iterate over the neurons in the previous layer.

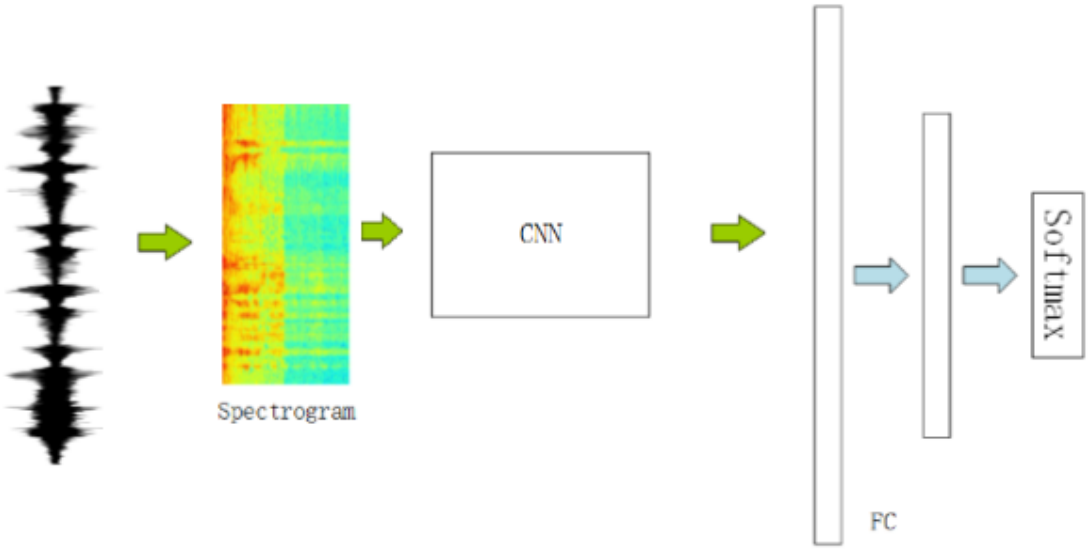

Figure 1.4 CNN architecture for spectrograms

Application to Speaker Recognition:

In speaker recognition, the input fed to the CNN is typically a spectrogram *S(f,t),* where *f* represents frequency and *t* represents time. The network learns filters and weights that capture relevant features for distinguishing between speakers. The overall process involves:

a. Convolutional Layers: Extract local features from the spectrogram
b. Hierarchical Feature Learning: As the signal passes through deeper layers, the CNN learns hierarchical features, recognizing more abstract representations of the speaker's vocal sound. This enables the model to discern unique characteristics and nuances
c. Pooling Layers: Downsample and reduce spatial dimensions
d. Flatten: Transform the 2D feature maps into a 1D vector
e. Fully Connected Layers: Learn high-level representations for speaker identification.

Advantages of CNNs in Speaker Recognition:

a. Local Feature Learning: CNNs excel at capturing local patterns in data, making them well-suited for tasks where local features are essential, such as identifying distinctive characteristics in a speaker's voice.
b. Automatic Feature Learning: Convolutional neural networks (CNNs) minimise the need for human feature engineers by automatically learning key features from data.
c. Robust to Variability: CNNs can handle variations in speech signals caused by factors like accent, language, or emotional state, making them robust for speaker recognition across diverse datasets.

Convolutional Neural Networks are powerful tools for speaker recognition when adapted to process spectrograms of speech signals. Their ability to automatically learn hierarchical features makes them effective in capturing the unique characteristics of a speaker's voice, contributing to accurate speaker verification and identification systems.

1.2.3 Performance evaluations

The evaluation of performance is the last phase in the creation of a standard structure of a speaker recognition system. Performance is often evaluated founded on the occurrence of errors. This section presents comprehensive information on the error and assessment metrics used in a standard speaker-recognition system.

1.2.3.1 Types of Errors

There are two main types of errors that might happen while evaluating a speaker-recognition system:

a. False acceptance
b. False rejection

False acceptance errors arise during a speaker-verification assessment when the system mistakenly identifies an impostor's speech as that of the target speaker.

The definition of the False Acceptance Rate (FAR) is as follows:

$$FAR = \frac{Total\ number\ of\ FA\ errors}{Total\ number\ of\ imposter\ speaker\ attempts} \quad (1.6)$$

Another form of mistake that occurs is known as false rejection, which occurs when the system incorrectly rejects the speaker who is the object of the communication.

The False Rejection Rate is defined as:

$$FRR = \frac{Total\ number\ of\ FR\ errors}{Total\ number\ of\ enrolled\ speaker\ attempts} \quad (1.7)$$

1.2.3.2 Performance metrics

The evaluation of performance often involves the use of two performance metrics:

(a) Decision cost function (DCF)

(b) Equal error rate (EER).

Decision Cost Function (DCF) The Decision Cost Function (DCF) is determined by attributing costs to both false rejection and false alarm mistakes while also considering the previous likelihood of client and imposter trials.

$$DCF = C_{miss}P_{miss}P_{target} + C_{fa}P_{fa}P_{imposter} \quad (1.8)$$

where,

$$C_{miss} = Cost\ of\ missed\ detection,$$

$$C_{fa} = Cost\ of\ false\ alarm,$$

$$P_{target} = Prior\ probability\ of\ the\ target\ trails,$$

$$P_{miss} = Percentage\ of\ the\ missed\ trails,$$

$$P_{imposter} = Prior\ prabability\ of\ the\ imposter\ trails,$$

$$P_{fa} = Percentage\ of\ the\ falsely\ accepted\ imposter\ trails.$$

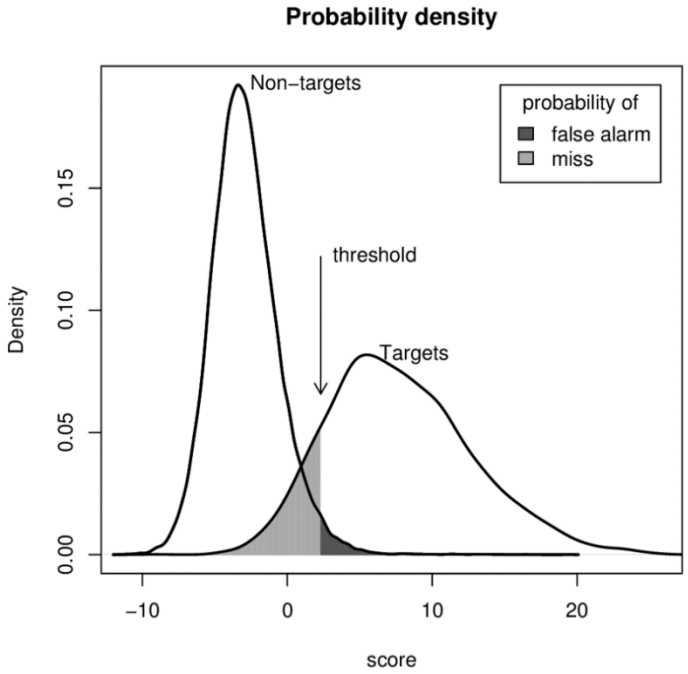

Figure 1.5 The score distributions for non-target (left) and target (right) trials [40]

Another kind of performance measure metric produced from DCF is the minimum DCF (minDCF). This metric represents the minimal value of DCF that is computed by adjusting the threshold.

$$minDCF = min[C_{miss}P_{miss}P_{target} + C_{fa}P_{fa}P_{imposter}] \quad (1.9)$$

The Equal Error Rate (EER) The Equal Error Rate (EER) is determined by modifying the threshold until the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) are equal. Figure 1.4 illustrates an instance of an EER selection process. The solid line represents the modified threshold at which the regions under the FRR and FAR are equivalent. While this measure is generally acknowledged, speaker-verification systems that rely on the Equal Error Rate (EER) threshold are not ideal for practical use. In order to minimise the false acceptance mistake, it is advisable to set a higher threshold for high-security applications.

1.2.4. Speaker segmentation and clustering outline

Speaker segmentation and clustering are crucial tasks in various fields, such as speech recognition, speaker diarization, and audio data indexing [41] [42]. It allows for the identification of different speakers in audio recordings and groups them based on their speech segments. Partitioning a speech signal into a sequence of speaker-homogeneous regions is the objective of a speaker segmentation and clustering system. Therefore, the answer to the query "Who spoke when?" may be found in the output of such a system, which is generally termed speaker diarization. This technology has evolved over the years, with improvements in accuracy and robustness. Speaker diarization is a vital component in numerous applications, including speech recognition and audio data indexing. By accurately separating speakers in audio recordings and clustering their speech segments, speaker segmentation and clustering enable tasks such as transcribing speeches, analyzing conversations, and organizing audio data based on speaker identities.

Segmentation is either presumed to be known or is carried out automatically before clustering. Nevertheless, methods that involve segmentation and clustering in a sequential manner have some restrictions [43] [44]. In actual applications, it is uncommon to have previous knowledge about the right segmentation in the former scenario. In the second scenario, any mistakes made during the process of dividing the data into segments might negatively impact the effectiveness of the next phase of grouping similar elements together.

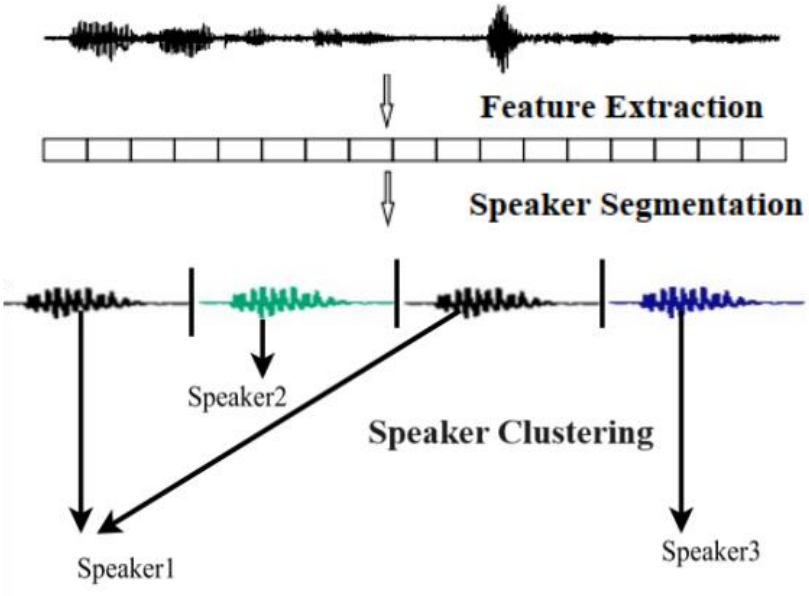

Figure 1.6 Speaker diarization system flow

Figure 1.6 depicts the standard process of a speaker diarization system. The process typically consists of two essential stages: Speaker Segmentation and Speaker Clustering. The audio stream is partitioned into comparable segments in terms of speaker identification, ambient state, and channel condition after the feature vectors have been extracted from the stream. The next step is to cluster the speech segments according to whether or not they come from the same speaker. Clustering and segmentation are usually done one after the other. Another option is to combine the segmentation and clustering processes in an interactive fashion, including clustering input into the segmentation process. These two phases can be repeated several times in an iterative manner.

1.2.5 Challenges in Speaker Recognition

The literature study on speaker identification highlights various challenges. Modern speaker recognition has issues such as managing background noise and speech quality fluctuations, handling numerous languages or accents, and resolving privacy and ethical considerations around the acquisition and use of voice data. In order to tackle these difficulties, scientists have devised many methodologies, including algorithms for reducing noise, methods for extracting robust features, and models particular to each language. Furthermore, the progress made in deep learning models, namely Convolutional Neural Networks have shown promise in improving speaker recognition systems via the automatic extraction of unique traits from raw audio data. The progress made in speaker identification, speaker verification, and speaker diarization has promise for improving a range of applications, such as voice-controlled devices, forensic analysis, security systems, and personalised services. One major obstacle is the need to identify specific features in speech signals for superior performance across various challenging elements. The other key challenges are:

a. The issue of channel mismatch arises when the training dataset uses one recording device, and the test audio involves a different channel, requiring an assessment of feature-to-channel mismatch performance.
b. Session variability, stemming from differences in recordings of the same speaker due to factors like changing settings and intra-speaker variability (mood, health, ageing, stress), poses a formidable challenge in speaker recognition.
c. Imposter speech, involving deceptive attempts to fool speaker identification systems, is a growing concern, encompassing both human and machine-generated spoofing and impersonation.

While speaker recognition systems primarily use speech for identification, the field is rapidly expanding for person authentication and security enhancement in various applications such as financial transactions, telephone shopping, voicemail browsing, and more. Despite its potential, the complexity of identifying suitable sound parameters makes designing reliable speaker identification systems challenging. Ongoing efforts are crucial to improving assessment methodologies for voice and speaker recognition systems. Although several research studies have been conducted, a comprehensive solution for real-time deployment of speaker identification technology is currently lacking, emphasizing the need for continued research and development in the field.

1.3 Background and related work

Recognising speakers has been extensively explored using a plethora of learning methods. The prevailing method for text-independent verification relies on Gaussian Mixture Models (GMMs) [34] [45]. The learning approach used is generative, where the input from each speaker is represented by a probability distribution. Specifically, a mixture of Gaussians is fitted to the data using the Expectation Maximisation (EM) technique [46]. Recently, researchers have explored the use of a Support Vector Machine (SVM) either as an independent module or in conjunction with a Gaussian Mixture Model (GMM) to enhance accuracy [47] [48]. Supervectors have emerged as a current study trend. A supervector is a fixed-dimensional representation of an utterance that effectively addresses the challenge of expressing utterances with varying lengths. The GMM-UBM, which stands for Gaussian Mixture Model-Universal Background Model, is used to associate an utterance with the supervector by using Maximum A Posteriori (MAP) adaption of the means [49]. The combination of supervectors with SVM has shown strong performance [50]. Recent research has focused on exploring factor analysis methods applied to GMM super-vectors. The goal is to separate the speaker-dependent components, which are crucial for speaker verification, from the channel and noise characteristics. Speaker verification was the first use of the i-vector method, which is based on joint factor analysis mentioned in [51]. It employs a complete variability transformation to convert the utterances of each speaker into a low-dimensional vector. Each speaker is represented by a singular vector that serves as a template for scoring based on cosine distance. The i-vector method is currently considered the most advanced technique for text-independent speaker verification [52] [53]. It is commonly used in combination with channel compensation methods such as Within-Class Covariance Normalisation (WCCN), Linear Discriminative

Analysis (LDA), and Probabilistic Linear Discriminative Analysis (PLDA) [54]. Artificial neural networks (ANNs) have been successfully used in many pattern recognition applications [55]. In speaker identification, specifically, its application aims to acquire characteristics that can withstand changes in communication channels [56]. In addition, they are under investigation for speaker verification, often in combination with an i-vector system [57]. In this case, a DNN that has been trained for voice recognition is used to replace the GMM in extracting the necessary statistics for i-vector extraction. A discriminatively trained deep neural network (DNN) may serve as an extractor of deep bottleneck characteristics. These features can be used for speaker verification, as shown in reference [58].

An extensive literature review is conducted, especially on existing approaches for speaker identification and speaker diarization, as well as the development of novel techniques needed for forensic applications. The literature examination of this study endeavour is segmented into four distinct sections. Table 1.1 provides information on these stages, including their goals and results. Various modern techniques and concepts are relevant to the region of research. The resume of the relevant literature is as follows:

With the advent of the voice spectrogram in the 1950s, a method for identifying speakers was introduced, thereby beginning the area of Speaker Recognition. The impact of several parameters on identification was investigated by J. Pollack *et al.* [59]. Features of voiced and unvoiced speech, the length and frequency range of the speech signal, the class size of potential voices, and the simultaneous display of multiple voices were some of the aspects taken into consideration in the research. The most important factor among them was found to be the length of the speech signal.

In their work, S. Pruzansky [60] used a pattern-matching algorithm to automatically identify speakers and investigate how variations in patterns impact identification accuracy. A time-frequency energy pattern was derived from the spoken words often spoken by ten individuals. The reference patterns were cross-correlated with the test pattern, and the pattern that exhibited the greatest correlation was selected. The research achieved a recognition score of 89%. Voiceprints are spectrograms of speech that are akin to fingerprints. Due to the unavailability of spectrogram data in real-time, the implementation of this technology was hindered, preventing the identification of enormous populations.

Table 1.1: **Different phases of literature review**

Stage	Area	Objective	Outcome
I	Speaker recognition system	To review and understand the concept and methodologies of the Speaker recognition system.	The effectiveness of the speaker recognition system is determined by feature extraction and feature modelling approaches.
II	Feature extraction	The objective is to analyse several feature extraction techniques and identify the most effective and reliable features for the speaker recognition system.	Various aspects, extracting methods, and assessing the effectiveness of different features and techniques are analysed.
III	Modelling	To review and understand various modelling techniques – generative and discriminative.	Various aspects and assessing the effectiveness of different modelling techniques are studied.
IV	Speaker Identification	To understand insights into various challenges and research areas within the field.	The primary outcome of the survey is the identification of key obstacles in the foreseeable future, particularly the need to determine specific characteristics in speech signals that yield superior performance across diverse challenges like forensic applications.
V	Speaker Verification	To get insights into numerous issues and study topics related to accuracy and robustness.	The study underscores the challenge of ascertaining specific characteristics in speech signals that can enhance performance across various demanding conditions, serving as a primary focus for different applications.
VI	Speaker Diarization	To get a comprehensive understanding of the intricacies and research domains.	The study highlights challenges associated with accurately segmenting audio recordings into speaker-specific segments. Factors such as background noise, overlapping speech, and variations in speech characteristics pose significant hurdles in achieving precise segmentation.

The selection of feature extraction strategies has a substantial impact on the performance and resilience of speaker recognition systems. Researchers have investigated many feature extraction strategies to improve the efficiency and reliability of speaker-related tasks. Each method has distinct strengths, limitations, and applicability to certain situations. This research explores the many methods used to extract features in speaker identification, verification, and diarization. The main objective is to examine and compare different methods, providing insight

into their effectiveness in dealing with issues including channel mismatch, session unpredictability, and impostor speech. By examining the intricacies of different feature extraction methods, this study aims to provide insights into the trade-offs, advancements, and potential breakthroughs that can impact the development of robust and accurate speaker-related systems. Table 1.2 provides a comparative analysis of different features. This comparative analysis aims to contribute valuable understandings into the strengths and limitations of diverse feature extraction techniques.

Table 1.2 Comparison of different features

Technique	Merits	De-merits
LPC	Derived from fundamental principles of sound generation	Noise-induced performance deterioration
	Easy to execute and mathematically accurate	More computational cost and time, It fails to accurately depict the vocal tract features resulting from glottal dynamics
LPCC	The spectral envelope is more even and the representation is more consistent compared to LPC	Insufficient linearly spaced frequency bands
	The components of the features are not connected with each other	Quantisation noise sensitive; insufficient order degrades performance
MFCC	Greater amount of information is available on lower frequencies compared to higher frequencies, as a result of mel-spaced filter banks. Therefore, it exhibits similar behaviour to the human ear when compared to other approaches	Fixed time-frequency resolution, because they are based upon STFT
	Summarises the key attributes of phones in speeches with little intricacy	Low robustness to noise

In this comparison, conventional methods of signal processing are considered. Features based on energy, formants, and spectral information, such as Mel-frequency cepstral coefficients (MFCCs), may be included of more conventional approaches. However, contemporary methods often use deep learning-based strategies, which have shown potential in detecting intricate patterns in voice signals. These approaches include convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention mechanisms. Utilising Gaussian mixture speaker models, D. Reynolds [61] presents state-of-the-art speaker identification and

verification techniques. These models successfully express speaker identification in a consistent and statistically generated manner, leading to impressive performance results. In contrast to the verification system's likelihood ratio hypothesis tester with background speaker normalisation, the identification technique uses a maximum likelihood classifier. The TIMIT, Switchboard, NTIMIT, and YOHO speech databases are used to test the systems.

In their study, D. Reynolds *et al.* [62] explore the issue of identifying and verifying speakers in noisy environments. They assume that speech signals are affected by ambient noise but do not have information on the specific properties of the noise.

In an audio recording with an unknown quantity of speech and an unknown number of speakers, the job of "who spoke when?" is known as speaker dialization. A variety of speaker diarization techniques were examined by Miro X. *et al.* [63].

In their study, Chaudhry G. *et al.* [64] examined feature extraction approaches both with and without noise correction strategies. The researchers determined that the Fourier transform, Auditory transform, and Hilbert transform are beneficial in various feature extraction techniques. Additionally, the characteristics of these features are influenced by the source of their study. The outcome of a technique is influenced by several factors, including restrictions such as the existence of noise. For instance, relying only on MFCC may provide excellent results in noise-free environments. However, when faced with mismatched testing and training settings or a larger population of speakers, it can lead to a decline in performance. Additionally, the combination of the feature extraction technique and matching model may vary from the feature extraction approach used for classification.

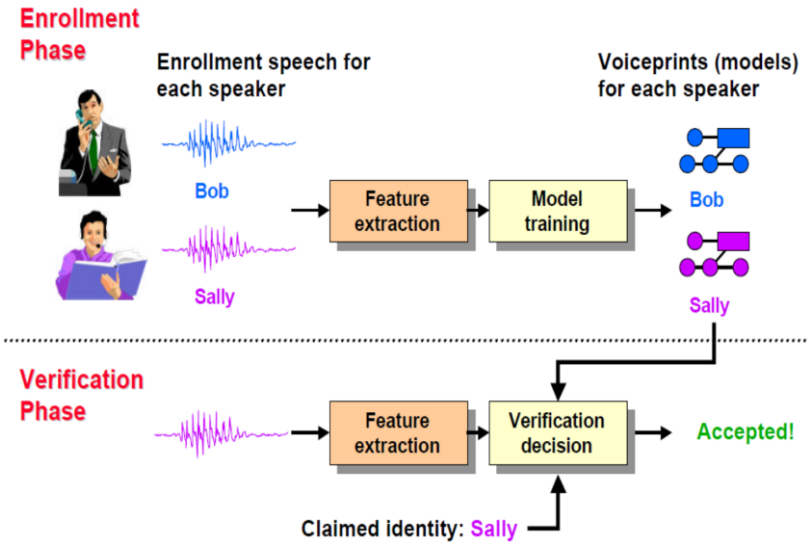

Figure 1.7 Phases of Speaker Verification [62]

Speaker identification is identifying a speaker based on their unique vocal features [65]. There are several speaker identification approaches [66] [67] [68] [69] [70], each with advantages and limitations [71]. One approach uses human listening, where individuals listen to voice samples and manually identify the speaker. The need to generate speech that sounds natural from speech processing equipment inspired the usage of this technology, which has been around since the beginning of speaker identification research [72]. However, this method has limitations such as subjectivity and lack of scientific basis in practical conditions. Another approach to speaker identification is the use of voiceprint analysis. Voiceprint analysis involves extracting acoustic features from a vocal sample and comparing them to stored voice models. Voiceprints represent the unique characteristics of an individual's voice, including pitch, tone, and vocal tract resonances [65]. This approach has been widely used in forensic speaker identification, where the goal is to match a suspect's voice with a recorded voice sample. One line of research in speaker identification focuses on evaluating performance in different environmental noise conditions [73]. These studies aim to understand how speaker identification systems perform in real-world scenarios where background noise or other environmental factors affect speaker identification accuracy. Another line of research focuses on the development of automated speaker identification systems [68]. These systems aim to automate the process of speaker identification, reducing the need for human intervention. These automated systems use advanced algorithms and machine-learning techniques to extract and analyse vocal features from a speech sample. Numerous semiautomatic speaker identification systems have been developed in the literature, each with advantages and limitations. Additionally, there has been significant research in improving the accuracy of speaker identification systems [34]. Some issues specific to forensic speaker recognition include the need to analyse speech samples obtained under realistic conditions, such as anonymous calls or wiretapping processes by police agents. Issues with speaker identification systems, such as text-independent recognition, have also been the subject of study [74].

Han *et al.* [75] have presented that the Speaker recognition system realizes better efficiency in controlled conditions. The occurrence of background noise was a major factor in distorting efficiency. Despite main advances in the speaker recognition field, noise and the restriction of obtainable speech data remain open issues, and the optimal solution to address them has not yet been found. Rubanenko *et al.* [76] presented a strategy to solve the challenge of optimising the primary electrical network. The presented system utilizes diminished gamma pitch coefficients for enhancing robustness through the limited duration of voice data. To show the

helpfulness of these coefficients compared with known characteristics, speakers use dissimilar database evidence in dissimilar conditions.

Venkatesan and Ganesh [77] have introduced the auditory scene analyzer, a system that uses speaker detection and binaural speech segregation, two continuous modules. We compared the source/artefact, source/distortion, and source/interference ratios to prior models to confirm the efficacy of network-based recurrent deep speech segregation. In addition, the mentioned system recommends the Gabor-Hilbert envelope coefficients and the spectra-temporal extractor. Several noisy and reverberant settings were used to verify the efficiency, and the outcomes were compared to preexisting monaural features. Despite a longer noise duration of 0.89 s compared to previous reference techniques, the findings of binaural speech segregation demonstrate that a realistic SNR of 0.7 dB is still achievable.

Sangeetha and Jayasankar [78] have introduced an extreme learning machine (ELM) based novel method for whispered speaker identification. When it came to applications for automatic speaker detection, the provided system was among the most challenging. Based on the variance between whispered and neutral speech under acoustic parameters, the effectiveness of typical speaker identification for neutral speech was much lower than for hushed speech. We compared the HPNP-ELM, a hybrid parametric and nonparametric probability density ELM evaluation, to the traditional ELM evaluation that relies on both types of data. The trial outcome shows that the proposed speaker detection system for whispers has made great strides.

Sun *et al.* [79] proposed that eliminating features like deep neural networks (DNNs) and Gaussian mixture models (GMMs) significantly enhances speaker identification. The new deep Gaussian correlation super vector (DGCS) characteristic depends on the DBN-GMM hybrid model to remove speakers' personality features with greater precision. Initially, the MFCC was removed as preprocessed speech signals, and then the DBN was used to gain bottleneck characteristics. Also, assuming the relevance among DGS deep means vectors, DGS was changed into DGCS via a super vector recombination process. The experimental outcomes portray that the use of DGCS may considerably recover a recognition rate of 17.979%, similar to a super vector system.

Shon *et al.* [80] have suggested frame-level speaker embeds for independent speaker identification, including end-to-end mode text. For understanding, the speaker identification model works with independent text input, and the framework was modified to remove the frame-level speaker embedding as every hidden layer. Particularly, the frame-level inlays owned to similar phonetic groups were equal (depending upon cosine distance) to the similar

speaker. The representation of frame-level permits for analyzing the networks under frame level and the possibility for other analyses to recover speaker recognition.

The Deep CNNs, which Kwon [81] has proposed, pay special attention to speaker recognition. Although deep learning has made great strides in recent years, most present-day speaker identification applications still depend on i-vector approaches. A method for speaker division that does not rely on text for recognition was suggested: deep convolutional NN. Two prominent convolutional neural networks (NNs)—the visual geometry group network and the residual neural network—form the basis of the approach. The deep convolution NN used the structured self-attention layer via many attention hops, but it couldn't handle segments of varying lengths. The suggested method surpasses both conventional i-vector-base models and CNN's additional strong baselines, according to the experimental findings conducted using the speaker recognition reference database, VoxCeleb. In addition, the results suggest that clustering unknown speakers using the higher layer of the embedded vector's previously trained recognition convolution NN was likely.

Nassif *et al.* [82] have given a pre-processing module dependent on Computational Auditory Scene Analysis (CASA) for the decreasing of noise as well as a "cascaded Gaussian Mixture Model – Convolutional Neural Network (GMMCNN) classifier to speaker identification" followed by emotion detection. To improve speaker recognition accuracy in emotionally charged and loud environments, a novel method was introduced. Results comparing the offered method to SUSAS, ESD, and RAVDESS show that it produces workable outputs.

Qaderi *et al.* [83] have introduced a novel combination of MFCC and time-based features (MFCCT), which enhances text-independent speaker identification accuracy by combining the efficacy of MFCC with time-domain information. The speaker identification was created by feeding the retrieved characteristics of MFCCT into a DNN. The findings reveal that the given MFCCT features coupled to DNN beat the prior baseline MFCC with the characteristics of time-domain features in the Libri-Speech dataset. In addition, when compared to five machine learning methods that were previously used for speaker identification, DNN achieves better categorization results. Both 1 and 2-level categorization modes are tested in this study to see whether one is better at speaker identification. The findings show that compared to the 1-level classification, the 2-level classification yields more practical results.

Shahin *et al.* [84] have put out a model that can better identify speakers when they're expressing strong emotions. The cascaded Gaussian Mixture Model-Deep Neural Network (GMM-DNN) classifier is the backbone of this model. Presentation, activation, and evaluation are the three

main focuses of the given model. The results demonstrate that speaker recognition in different moods is improved by the cascaded GMM-DNN classifier. Using cascaded GMMDNN, we were able to accomplish speaker identification presentation, which was identical to the results we got from human listeners' subjective evaluations.

Some recent studies *on Speaker verification* have focused on feature extraction approaches, which greatly impact the performance of the system [85] [86] [87] [88]. These studies have investigated and assessed the efficacy of several feature extraction methods for speaker verification tasks, including i-vector quantization, linear predictive coding, and Mel frequency cepstral coefficients. Evaluating optimisation strategies used in speaker verification is another crucial component of the literature study. The application of complex systems makes use of multi-step optimisation methods for complex dynamic systems, which may include computer, mathematical, analogue, full-scale, and semi-natural approaches [89]. These techniques aim to improve the efficiency and accuracy of speaker verification systems by optimizing various components, such as feature extraction, model training, and threshold selection [90] [91]. Additionally, model validation is an essential aspect of speaker verification systems [45] [92]. It involves testing the performance and generalization abilities of the competent model on unseen data. Some common approaches to model validation in speaker verification systems include cross-validation, where the data is alienated into multiple subsets for training and testing, and leave-one-out validation, where one data point is left out for testing while the rest is used for training [32] [93].

Hu, Q. *et al.* [94] as a resource for learning about the latest developments and trends in speaker verification, provide an evaluation of the technology for the state-of-the-art biometrics excellence roadmap. By including the Mel-frequency cepstral coefficients (MFCC) derived from the Discrete Wavelet Transform (DWT) model, Al-Ali *et al.* [95] aims to improve the efficiency of speaker verification, especially when dealing with reverberation and noise. Huang *et al.* [96] provide a solution to problems with Fuzzy C-Means (FCM) and an enhanced natural exponential inertia weight Particle Swarm Optimization (IEPSO) technique for feature extraction, pre-classification, and classification utilising a multiple-layer hybrid fuzzy support vector machine (MLHF-SVM). Because it outperformed competing algorithms, Kalam *et al.* [97] used the PSO algorithm to solve the optimization challenge.

Swain *et al.* [98] developed a Deep Neural Network (DNN) method founded on a 2D-Convolutional Neural Network (CNN) and Gated Recurring Unit (GRU) for speaker recognition, utilizing convolution layers for dimensionality reduction and voiceprint feature

extraction. Mardhotillah *et al.* [47] employ a Support Vector Machine (SVM) model for recognizing telephone conversations compared to unexpected sound recordings. Saleem *et al.* [99] propose a novel forensic speaker recognition methodology involving language and accent data extraction from short words, utilizing CNN-based models like GMM-CNN and VGGVox, as well as the x-vector model founded on Deep Neural Networks (DNN).

Devi *et al.* [100] present a hybrid mechanism for Automatic Speech Recognition (ASR) with an Artificial Neural Network (ANN) model, employing MFCC for feature extraction, Self Organizing Feature Map (SOFM) for dimension reduction, and a Multilayer Perceptron (MLP) model with Bayesian Regularization for recognition. Teixeira *et al.* [101] demonstrate the extraction of speaker embedding while keeping the speaker voice and service provider model private, using Secure Multiparty Computation, aiming for a balance between computation and security costs in Automatic Speech Recognition (ASR).

Speaker Diarization systems aim to automatically label audio or video recordings with speaker identities, allowing for efficient analysis and organization of speech data. The techniques used in speaker diarization include traditional statistical modelling methods, clustering algorithms, and machine learning techniques. Park *et al.* [102] discuss the integration of speaker diarization systems with speech recognition uses and the potential for deep learning to enhance the performance and efficiency of these systems. It can be observed that speaker diarization technologies have evolved from being a mere preprocessing step for speech recognition to becoming an important and integral component of various applications such as conversational AI, audio indexing, and audiovisual diarization. Based on the review conducted by Anguera *et al.* [41], it is evident that speaker diarization technologies have undergone significant advancements in recent years. Deep learning-based techniques have shown encouraging results in boosting accuracy and resilience in speaker diarization tasks, which the paper emphasises as a departure from older methods. Additionally, the review emphasizes the increasing integration of speaker diarization systems with other applications, such as speech recognition and audio indexing, to enhance their performance and utility in real-world scenarios. The sources provide a comprehensive overview of speaker diarization research, including different methods and techniques used, recent advancements in deep learning approaches, and the integration of speaker diarization with other applications.

Khoma *et al.* [103] present two speaker detection system structures that rely on diarization integration. These systems operate based on either group-level or segment-level classification. The study includes four experiments aimed at selecting the optimal supervised diarization

technique using PyAnnote. The first experiment explores how the choice of distance function in vector embedding influences the reliability of speaker detection in the segmentation-level classifier structure. Kumar *et al.* [104] analyze the impact of contextual elements on diarization efficiency in communication, identifying factors influencing all types of diarization errors. To enhance performance, a Deep Neural Network (DNN) is trained on diarization output combined with features.

Prokopalo *et al.* [105] introduce a structure involving human interaction to correct clustering through modest queries. The author outlines the method for listing individual queries and two ending conditions to alleviate the workload on humans in the loop. In reference [106], a comprehensive analysis of the RPNSD technique is conducted with two significant contributions. Firstly, the diarization efficiency is reported on other databases, and the impact of various system settings is analytically examined. Secondly, an automated speech recognition (ASR) element is integrated into the RPNSD method, resulting in a novel structure named RPN-JOINT that concurrently performs diarization and ASR. The authors [107] assess the performance of this multimodal speaker diarization mechanism under noisy conditions, utilizing realistic environmental noise and additive white Gaussian noise (AWGN). It is suggested to incorporate an LSTM-based speech enhancement block in the diarization pipeline to mitigate the noise effect. Two enhanced approaches are introduced based on the self-attention system, eliminating the application of local correlation and seeking the same speaker embedding from entire sequences [108]. One method achieves existing performance on the DIHARD II Eval Set, while the second exhibits higher efficiency. Tran and Tsai [109] present two DL-OSID methods: single-stage and two-stage. The two-stage system identifies simultaneous speakers after initial speaker detection.

These studies collectively contribute to the evolving landscape of speaker identification, speaker verification and speaker diarization systems, introducing novel techniques and demonstrating improved performance across various challenging conditions and scenarios, but speaker recognition for forensic applications is still a challenge.

1.4 Research gaps

Research is a continuous process, evolving novel methods and concepts for the betterment of humanity. Researchers have made a lot of progress in speaker identification so far, but there is still a long way to go before things are perfect. After reviewing the available literature, the following scope has been found for the research.

Within the realm of forensics, the technique of speaker verification has the potential to be used in order to produce evidence conclusions in legal disputes. Unlike fingerprinting and DNA-based authentication technologies, the present speaker verification solutions are restricted by their sensitivity to noise interference and vulnerability to signal manipulation using voice recording equipment. This is in contrast to the technologies that are utilising DNA to authenticate individuals. When it comes to the development of technology for speaker verification and identification, the area of biometric applications has a great deal of potential. However, there are still considerable difficulties that need to be addressed in order to overcome them, and these obstacles are now preventing the broad use of the technology.

The issue of developing a robust speaker recognition system for unconstrained situations *(*forensic environment*)*, which is robust against background noise, channel mismatch, multi-speaker speech conditions, *etc*., is still unresolved. Speaker diarization is not yet another area where a lot of work is needed to be done. It is essential to produce more extensive datasets to strengthen the systems' ability to withstand undiscovered abnormalities and to increase the significance of the findings. With the increasing volume of datasets, it is unavoidable that systems would need to improve their efficiency to process such data within a reasonable timeframe. Assigning overlapping speech to several speakers remains the most challenging hurdle to solve.

1.5 Objectives

- To study and analyse the different available biometric and forensic speaker recognition models.
- To develop a speaker identification model using available databases for forensic application.
- To develop a model for speaker diarization for conversational speech.
- To develop a speaker verification model for conversational speech using various available databases for forensic application.

1.6 Applications

With the growing prevalence of mobile phones and the abundance of sensitive information they store, it is crucial to incorporate effective biometric authentication alongside the conventional password authentication system. The data and files saved on the device will be safe and secure because of this. Biometric systems have made heavy use of speaker recognition technologies in

recent years. It offers more advantages than other biometric methods like fingerprints, facial recognition, and iris scans, as shown in Figure 1.8.

Figure 1.8 Voice as biometric [110]

The prime and voice recognition systems exhibit commonalities in their pattern-matching procedures and may also be integrated. Speaker recognition and diarization are crucial technologies that have diverse applications across numerous fields. Speaker recognition in the realm of law enforcement may be used to ascertain the identity of suspects or validate the genuineness of recorded conversations during criminal investigations. Speaker recognition in the domain of customer service enables the customisation of contacts with clients by discerning their preferences and requirements via analysis of previous encounters [42]. Speaker recognition in healthcare may be used for patient identification and authentication, guaranteeing that only authorised persons can access confidential medical data. Speaker diarization is a technique in the media and entertainment industry that may be used to automatically produce subtitles or captions for videos, hence enhancing accessibility for those with hearing disabilities. Furthermore, speaker diarization may be used in research and academia to analyse extensive collections of audio recordings, such as lectures or interviews, in order to extract significant insights and identify patterns. Several fields find various applications for speaker identification and diarization systems, including healthcare, research, media and entertainment, customer service, and law enforcement [41]. According to Park *et al.* [42], cloud-based voice recognition technologies improve several aspects of technology, such as web browsing, transportation, and healthcare. For instance, using voice commands enables drivers to browse the Internet without compromising road safety hazards. Network congestion may significantly impact the usability

of these apps, leading to user displeasure. When this occurs, cloud-based voice recognition systems may benefit from speaker detection and diarization technologies, which would make them more efficient and easier to use. These technologies properly identify and separate speakers in audio recordings, leading to more precise and personalised interactions. Additionally, speech recognition and diarization may be used in voice assistants and smart home devices. Speaker recognition may be used to distinguish between various users and provide personalised replies and assignments. The implementation of this degree of customisation elevates the user's satisfaction and optimises the proficiency and efficacy of voice-activated devices in executing functions like scheduling reminders, playing music, and managing smart home gadgets. Speaker recognition and diarization have diverse applications in several sectors like customer service, healthcare, media and entertainment, law enforcement, research and academia, and voice assistants/smart home devices [42]. To summarise, speaker recognition and diarization systems have extensive applicability across several sectors. Additionally, they enable the analysis and extraction of insights from large datasets of audio recordings. Moreover, they can personalise interactions with voice assistants and smart home devices.

1.7 Organisation of the Thesis

There are five sections to the thesis. This section provides an overview of the thesis's structure and contents.

Chapter 2 In this chapter, the speaker identification model is discussed in detail. This chapter describes the methodology, discussions, and analysis of results used to develop the speaker identification model compared with other available techniques.

Chapter 3 In this chapter, the speaker verification model is discussed in detail. This chapter describes the methodology, discussions, and analysis of results used to develop the speaker verification model compared with other available techniques.

Chapter 4 This chapter discusses speaker segmentation and clustering. The system performance is demonstrated in different domains. In this chapter, the speaker diarization model is discussed in detail. This chapter describes the methodology, discussions, and analysis of results used to develop the speaker diarization model compared with other available techniques.

Chapter 5 wraps up the thesis and delves into various potential paths moving forward. Also covered are the potential practical uses of this work.

At the end, the **References** cited in this have been listed.

1.8 Contributions of the thesis

The variance in the signal characteristics is the most important aspect that affects the performance of speaker recognition. Variations may be attributed to the speakers themselves as well as to the recording and transmission channels. Some examples of these causes include:

- Variations that occur over a short period of time owing to the speaker's health and emotions
- Variations that occur in different environments
- Alterations that occur over time as a result of ageing
- Variable microphones

The speaker and channel effects are interconnected throughout the spectrum; hence, the attributes used in speaker recognition systems include both speaker and channel properties. This is the underlying cause behind this situation. Hence, every factor that influences the spectrum has the capacity to generate complications in speaker recognition. As a consequence of the speech being corrupted, the spectral-based characteristics are also impacted, leading to changes in the distributions of these features. Consequently, a speaker model trained on audio from a certain kind of corrupted environment would often exhibit poor performance in recognising the same speaker using data collected in other situations. The reason for this is the shift in feature distributions. When testing with damaged or distorted voices, the results demonstrated a significant drop in the accuracy of speaker recognition systems. The use of combination strategies to mitigate the problems caused by acoustic mismatches in speaker detection is proposed. Additionally, technologies that provide reliable speaker recognition are developed and deployed. A research on the process of speaker segmentation and clustering is conducted and a system with the objective of achieving high performance in situations involving several speakers while also being adaptable to other domains is developed.

This thesis focuses on conducting research aimed at enhancing the resilience of speaker recognition systems. The main objective of the thesis is to improve efficient characteristics for speaker recognition in order to get exceptional results in uncontrolled (forensic) environments. This will facilitate the use of the model in forensic applications. This thesis presents unique algorithms for robust speaker identification and verification, which tackle issues arising from differences in speech signals caused by varied forensic settings, such as environmental conditions, recording equipment, and channel distortions. Speaker recognition techniques are

made far more accurate and resilient by using state-of-the-art approaches for generating models and extracting unique properties. A comparison study was undertaken to assess the system using different attributes to showcase the efficacy of the suggested features and modelling methodologies. The thesis also suggests thorough assessment criteria for evaluating the effectiveness of speaker identification, verification, and diarization systems. Conducting comparative research against current methodologies allows us to obtain a deeper understanding of the strengths and limits of the suggested approaches.

The thesis might make a difference by outlining thorough criteria for judging how well speaker identification, verification, and diarization systems work. Insights into the strengths and limitations of the recommended methodology may be obtained by performing comparison studies with the existing methods. Conclusively, a thesis focused on robust speaker identification, verification, and diarization has the capacity to improve the current level of knowledge in this field by providing innovative algorithms, resolving practical challenges, and contributing to the overall progress of speaker-related technology. This is because the thesis will tackle the existing issues in the subject.

2. Chapter - Speaker Identification

The field of Speaker Identification, encompassing the discernment of an individual solely through their voice, has undergone a systematic evolution over the last few decades. The early stages involved manual assessments, combining aural and visual examinations facilitated by humans, notably through spectrograms. Subsequent phases saw advancements characterized by feature extraction and evaluation, culminating in contemporary methodologies reliant on Connectionist and/or Statistical models for automatic identification. A consistent trajectory in research during this temporal span has been directed towards establishing robust models that encapsulate speaker-dependent features resilient to variations induced by transmission and recording mediums. Speaker Identification, defined as the scientific process of recognizing an individual within a known population through voice analysis, is closely aligned with Speaker Verification [65]. The second one is checking a person's voice for signs of their claimed identity. Analytical and decision-making approaches based on a library of reference templates are common to both methods. A set of attributes derived from audio samples that are specific to each speaker is called a template in this setting. In the training phase, these templates are created. In the test phase, they are used for comparisons. It should be mentioned that the database size in the speaker identification studies only includes a predetermined group of N speakers.

The two possible outcomes of Speaker Verification—acceptance or rejection—make it a binary decision-making process. Typically, a threshold is established to facilitate this decision-making [111]. However, emphasis in this chapter is directed towards the domain of Speaker Identification. It is possible to think of the verification task as an open-set scenario within the identification task, with the variable N reduced to a single entity during the test phase since the identity is declared. In forensic science, speaker identification is of the utmost importance, since it uses audio analysis to identify individuals. Conversely, verification finds relevance in the security industry, where the validation of claimed identity through voice analysis is integral to ensuring secure access and authentication. Speaker Identification may be classified into two categories: Text-Dependent and Text-Independent. These categories share common signal-processing approaches [112]. The differences between these identification techniques mostly appear in the combinations of extracted characteristics used and the temporal structure that governs the identification process. It has been successful to use both short-term and long-term studies to traits that are statistically based and that are dynamic. Nevertheless, it is important

to highlight that text-independent systems tend to prioritise long-term analysis and statistically based methods. Obtaining a voice sample from a forensic or unrestricted setting sometimes results in a decrease in the performance of the identifier. A speaker identification system must have tolerant features, appropriate pre- or post-filtering mechanisms, or a model that describes the characteristics of the noise or channel to be robust in the presence of distortion and noise. Studies have shown that using these principles may greatly enhance the performance of speaker identification systems, enabling them to function successfully even in difficult acoustic situations characterised by background noise and distortions [113]. This chapter aims to consolidate the most effective strategies in these frameworks and outline the developing a robust speaker identification system. Following a comprehensive examination of speech processing theory in the first chapter, this chapter delves into the modern approaches and procedures that are used in the process of speaker identification.

2.1 Speaker Identification Model Framework

Speaker identification is the process of using artificial intelligence (AI) to identify and distinguish human voices. Speaker identification technique is widely used in speech recognition, security, surveillance, electronic voice eavesdropping, and identity verification. The current approaches lack the necessary accuracy and robustness in processing the voice signal. A Speaker Identification framework is developed to address these challenges. It is built on a Mask region-based convolutional neural network (Mask R-CNN) classifier that is optimised via Hosted Cuckoo Optimisation (HCO). The objective of the proposed methodology is to improve the precision and reinforce the robustness of the signal. Initially, the input speech signals are extracted from the real-time dataset. Subsequently, four distinct types of features are derived from the input speech signal, namely Mel Frequency Differential Power Cepstral Coefficients (MFDPCC), Power Normalised Cepstral Coefficients (PNCC), Gamma tone Frequency Cepstral Coefficients (GFCC), and Spectral entropy. These features are employed to enhance the robustness of the signal. Furthermore, speaker identification is accomplished through the utilization of the Mask R-CNN classifier. Likewise, the parameters of the Mask R-CNN classifier are fine-tuned using the HCO method.

2.1.1 State of art

Because of the application of voice access and security demand, speaker identification is rapidly growing in the speaker recognition field today [114]. Greenberg *et al.* [115] divide speaker recognition into three sub-fields: Speaker Identification (SI), Speaker Verification

(SV), and speaker classification. The system attempts to identify the unknown person's voice among a group of people in speaker identification [116] [117]. In speaker verification, the system compares the confirmed speaker's voice to the claimed person's voice and then uses a threshold to decide whether to accept or reject the procedure. [118]. Speakers are categorised in speaker categorization based on attributes such as gender and age. There are two distinct perspectives on speaker recognition [119] [120] [121]; they are (i) text-dependent and (ii) text-independent. In 1st case, the system is concerned with voice characteristics.

Both the characteristics of the voice and the words themselves are attractive to the system in the second scenario [119, 122]. There are primarily three steps to the speaker identification system: extracting features, modelling, and scoring. The speech signal is dynamic because speech activities are in a continual state of flux [123] [124]. Here, the signal is divided into smaller frames with a fixed signal strength, and spectral feature vectors are used to describe each frame [119] [125]. Using sub-band energy values as characteristics, the fast Fourier transform (FFT) was an earlier feature [126]. Typically, the feature vectors' dimensionality is diminished by other transformations named Mel frequency cepstral coefficients (MFCC) that are broadly utilized in the literature for a variety of speech processing uses, like speaker identification, speaker classification, speaker verification, and speech recognition. Both the characteristics of the voice and the words themselves are attractive to the system in the second scenario [119] [122]. There are primarily three steps to the speaker identification system: extracting features, modelling, and scoring. The speech signal is dynamic because speech activities are in a continual state of flux [123] [124]. Here, the signal is divided into smaller frames with a fixed signal strength, and spectral feature vectors are used to describe each frame [119] [125]. Using sub-band energy values as characteristics, the fast Fourier transform (FFT) was an earlier feature [126]. Typically, the feature vectors' dimensionality is diminished by other transformations named Mel frequency cepstral coefficients (MFCC) that are broadly utilized in the literature for a variety of speech processing uses, like speaker identification, speaker classification, speaker verification, and speech recognition [127] [128] [129].

Their capacity to simulate the vocal tract's structure within a shorter time power spectrum is the source of MFCC power. Logarithmic compression using the discrete cosine transform (DCT) and a filter bank inspired by psychoacoustics were the usual tools they used for their calculations [130] [131]. Finally, to represent the Mel frequency cepstral coefficients feature vector, the 12–15 least DCT coefficients are brought to light. Line spectral frequencies (LSF), perceptual linear prediction (PLP), and linear predictive cepstral coefficients (LPCC) are

among the feature extractors used in speech processing [132]. In reality, MFCCs are challenging to master. This is where you may activate the feature enhancement or feature selection.

This work introduces an effective Speaker Identification framework, employing a Mask R-CNN classifier whose parameters are optimized using Hosted Cuckoo Optimization (HCO) to enhance accuracy and bolster signal robustness. Real-time input speech signals are processed through a dataset, from which four distinct features are extracted: Mel Frequency Differential Power Cepstral Coefficients (MFDPCC), Power Normalized Cepstral Coefficients (PNCC), Gamma tone Frequency Cepstral Coefficients (GFCC), and Spectral entropy. These features serve to enhance signal robustness. Subsequently, speaker identification is conducted using the Mask R-CNN classifier, with classifier parameters fine-tuned via the HCO algorithm.

2.1.2 Chapter contributions

Among this work's most significant contributions are,

- Enhancing the accuracy and robustness of the speaker's voice signal, we provide an Efficient Speaker Identification Framework that relies on the Mask R-CNN Parameter Optimised utilising Hosted Cuckoo Optimisation (HCO).
- In the beginning, the input speech signals are extracted from the real-time dataset.
- From the input speech signal, four types of features are extracted: Mel Frequency Differential Power Cepstral Coefficients (MFDPCC) [119], Gamma tone Frequency Cepstral Coefficients (GFCC) [133], Power Normalized Cepstral Coefficients (PNCC) [134] and Spectral entropy [135].
- The Mask R-CNN classifier is then used to categorise the speaker ID using the retrieved characteristics
- Mask R-CNN classifier has concealed any acceptance of optimization strategies for calculating optimal parameters to declare the correct classification of speaker ID.
- Hence, Hosted Cuckoo Optimization (HCO) is proposed for optimizing the weights parameters of the Mask R-CNN classifier.
- Finally, the optimum hyper-parameter is carefully chosen in the Mask R-CNN classifier with the help of Hosted Cuckoo Optimization (HCO) [136] [137].
- The MATLAB platform is used to accomplish the planned task. Criteria for evaluating the suggested method's performance include sensitivity, specificity, accuracy, recall, F-score, and precision.

- The proposed Mask-R-CNN-HCO method is evaluated in comparison to other methods that have been previously developed for speaker identification. These methods include Automatic Classification of speaker identification using the K-Nearest Neighbors algorithm(KNN) [138], classification of speaker identification utilizing a multi-class support vector machine (MCSVM) [139], classification of speaker identification using Gaussian Mixture Model – Convolutional Neural Network (GMMCNN) classifier [82], classification of speaker identification utilizing Deep neural network (DNN) [116], classification of speaker identification using Gaussian Mixture Model – deep Neural Network (GMMDNN) classifier [140].

2.2 An efficient Speaker Identification framework using Mask R-CNN-HCO

We suggest using Mask R-CNN-HCO to identify speakers. Figure 2.1 shows the schematic of the Mask R-CNN-HCO-based Speaker Identification system. In this instance, a real-time dataset is used to get the input speech signals. Four types of features are extracted from the input speech signal to improve signal robustness: MFDPCC, GFCC, PNCC, and Spectral entropy. After that, the speaker ID is classified using the Mask R-CNN classifier. The Mask R-CNN classifier parameters are then optimised using the HCO algorithm.

The existing methods do not provide sufficient accuracy and robustness of the speech signal. To overcome these issues, an efficient Speaker Identification framework based on the Mask region-based convolutional neural network (Mask R-CNN) classifier parameter optimized using Hosted Cuckoo Optimization (HCO) is developed.

2.2.1 Speaker identification framework and speech signal Acquisition

The speaker identification approach is the method that should be used in order to achieve reliable and high-quality speech recognition in audio sources. There is a method that is referred to as "speaker identification" that involves the use of artificial intelligence to detect human sounds. Some of the various uses of this technology include identity verification, electronic eavesdropping, monitoring, and voice authentication. Other applications include security and monitoring. For the purpose of extracting these vocal signals, a real-time dataset was used.

2.2.2 Feature extraction

Extraction of discriminative with salient features from speaker utterances is critical for precise speaker identification. The process of removing unwanted or unnecessary information from signals and obtaining the required features from the feature identification process is known as feature extraction. In this manner, four types of features, such as MFDPCC, GFCC, PNCC, and

Spectral entropy, are extracted to identify the speakers efficiently. Optimal features of these metrics are their quantifiable and extractable nature, resistance to noise and distortions, and independence from environmental factors like speaker health and background noise.

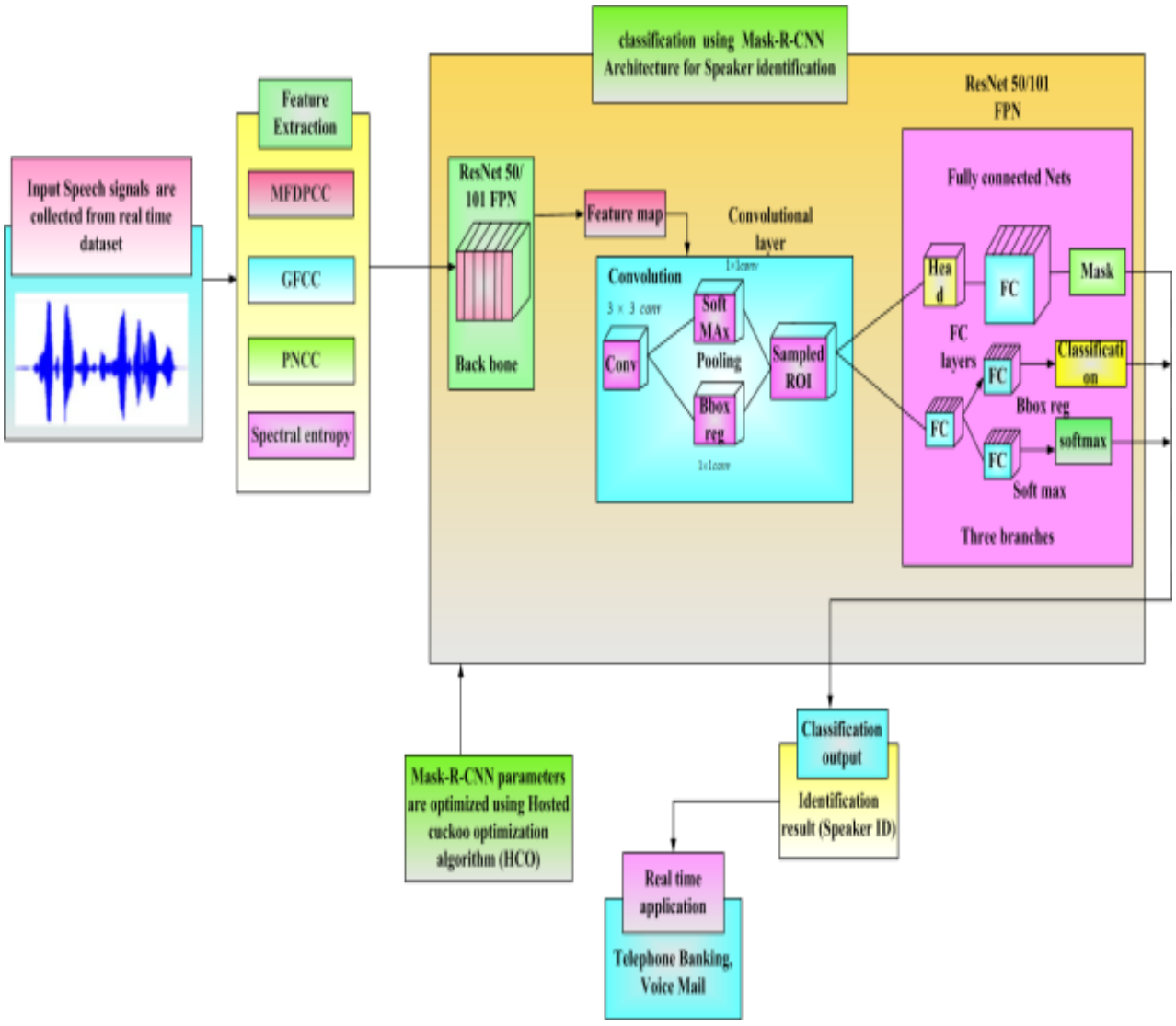

Figure 2.1 An efficient Speaker Identification framework using Mask-R-CNN-HCO

2.2.2.1 Mel Frequency Differential Power Cepstral Coefficients (MFDPCC)

The MFDPCC method is used to track the features of the spectral envelopes of the speech signals. This method is used to determine the voice and in the sound recognition process. In this case, the discrete cosine transform is used to obtain the voice signal's short-time power spectrum, D_c and the equation is given in equation 2.1

$$D_c = \sum_{\{L=1\}}^{20} log\ (log\ (F_L)).cos\left[m.\left(L-\frac{1}{2}\right)\right] \quad ,m = 1,2,\ldots.N \qquad (2.1)$$

From the equation (2.1), it is used to estimate the MFDPCC parameters. This F_l is represented as the energy in the appropriate $L-th$ filter with the $L-th$ sub-bands. From this, differentiated

Mel frequency is determined using the DFT signals. This is fixed using the Mel scale. Then, the Mel scale is measured using the Hertz scale. Mel scale equation is given in equation (2.2)

$$D[Scale_{mel}] = 2595.\log_{10}\left(1 + \frac{d(Hz)}{700}\right) \quad (2.2)$$

where, $d(Hz)$ is represented as the frequency of the sub-bands with the discrete Fourier transform. The input speech signal is received from the various speakers; the equation is given in equation (2.3)

$$h(f) = z(f) * g(f) + m(f) = a(f) + n(f) \quad (2.3)$$

where $h(f)$ is represented as the input speech signal, $z(f)$ is represented as the original clean speech signal, $g(f)$ is represented as the impulse response of the transmitting channel, $*$ is represented as the convolution operator, $n(f)$ represents ambient noise, $a(f)$ is represented as the noise-free speech signal. Then, the differential Mel frequency $d(f)_{mel}$ equation is given in equation (2.4):

$$d(f)_{mel} = H'(f) = \frac{dH(f)}{df} \quad (2.4)$$

Where the equation (2.4) is represented as the differentiation with respect to the f .In which $H'(f)$ and $H(f)$ is represented as a differentiation factor. Then, the non-correlation of the differential equation is given in equation (2.5)

$$d(f) = \frac{dH(f)}{df} = \frac{dA(f)}{df} + \frac{dN(f)}{df} = d_a(f) + d_n(f) \quad (2.5)$$

where, $d_a(f)$ and $d_n(f)$ is represented as the differential power spectra taken from the input noise-free speech and noise signal. Then, the MFDPCC equation is given in equation (2.6):

$$d(L)_{MFDPCC} \approx \sum_{k=-P}^{J} y_k H(L+k) \approx \sum_{k=-P}^{J} y_k \,[H(L+k) + N(l+k] = d_a(l) + d_n(l) \quad (2.6)$$

where P and J is represented as the differential order of the weighting MFDPCC coefficients with the value of $0 \le l < l$, l is represented as the FFT length. Where y_k is represented as the real value weighting coefficient, L is represented as the length, K and N is represented as the orders of the differential equations. In this way, the MFDPCC features have been extracted from the input dataset. It is also utilized to identify the speakers used in the telephone banking and voice mailing real-time applications.

2.2.2.2 Gammatone Frequency Cepstral Coefficients (GFCC)

The GFCC features are extracted from the input speaker. The filtering process is mostly used in forensic automatic speaker identification. This method is used to decrease the noise and to improve the robustness of the speech signal. This GFCC is determined by using the cochlea in the inner ear, which is mostly used in forensics. It is determined using the gamma tone with the center frequency ω_g , and the equation is given in equation (2.7):

$$h(f) = xt^{m-1}e^{-2\pi yt}\ cos\ cos\ (2\pi\omega_g + \varphi) \tag{2.7}$$

where t represents the time, ϕ represents the phase, and it is set as 0, x represents the constant that controls the gain, m represents the order of the filter, y represents the attenuation factor, $e^{-2\pi yt}$ represents the gamma function. Then the attenuation function y equation is given in equation (2.8):

$$y = 25.17\left(\frac{4.37\omega_g}{1000} + 1\right) \tag{2.8}$$

where ω_g is represented as the Gammatone filter with a centre frequency.

Then, the feature extracted GFC coefficient equation is given in equation (2.9)

$$f(g) = \sqrt{\frac{2}{M}}\sum_{g=0}^{M-1} F_g[n]\ cos\ cos\ \left(\frac{j\pi}{2M}(2g+1)\right),\ j = 0,1,\ldots,M-1 \tag{2.9}$$

Eq (2.9) is known as the feature-extracted GFCC method.

Where $f(g)$ is represented as the GFCC feature extracted equation. Where M is represented as the filter coefficient, F_g is represented as the time frequency, g is represented as the channel. Here, the cochlea gram is known as the time-frequency representation during the voice recording. In this way, the GFCC features have been extracted from the input dataset. It is also utilized to identify the speakers used in the telephone banking and voice mailing real-time applications.

2.2.2.3 Power Normalized Cepstral Coefficients (PNCC)

Here, The PNCC approach is used to enhance the signal's robustness via the extraction of speaker signal characteristics for speaker identification. These features are used to lessen the computational time, and the PNCC equation is given in equation (2.10):

$$J[n,k] = \sum_{l=1}^{N/2}|B[n,k]|^2\ H_k(L) \tag{2.10}$$

where B specifies attenuation factor, n specifies frame index, k is represented as the gamma tone filter index, B is represented as the Spectrum resolution, $H_k(L)$ is represented as the function of filter bank in the frequency domain. This method is employed to suppress the noise to increase the robustness [141]. This method is used to reduce channel bias, such as the speech signals being in the form of low and high frequencies. From this, channel bias noise is reduced by smoothening the speech signal. Then, the reducing channel bias $\overline{O}[k]$ equation is expressed in equation (2.11):

$$\bar{O}[k] = O[k] - f \times (O[n,k]) \tag{2.11}$$

Where, f is represented as the constant bias factor. Here, the auditory system consists of automatic gain control mechanisms, which reduce the amplitude effects of the received speech signals. In the PNCC method, the effect of the potential effect of the amplitude scaling is reduced by applying the mean power normalization. It is defined as the input power is normalized by dividing the received power in the running average by the total power. Then the PNCC with reduced mean power normalization $\eta[n]$ equation is given in (2.12):

$$\eta[n] = \delta_\eta \eta[n-1] + \frac{(1-\delta_\eta)}{l} \sum_{r=0}^{K-1} \bar{O}[n,k] \tag{2.12}$$

where δ_η is represented as the forgetting factor, its value is given as 0.999, l is represented as the arbitrary constant. Then, the normalized power is calculated with the help of running power $\eta[n]$, and the equation is given as:

$$U[n,k] = l \frac{\bar{O}[n,k]}{\eta[n]} \tag{2.13}$$

The speakers used in the real-time applications of telephone banking and voice mailing are identified using the PNCC characteristics that are retrieved from the input information.

2.2.2.4 Spectral entropy

A better signal-to-noise ratio may be achieved by using the spectral entropy. The signals are filtered using this method. Let us consider the input speech signals as $a_m(m = 1,2,....M)$, which include various kinds of noises, such as random noise, periodic, harmonic impact. The output of the signal is given as $D = (d_1, d_2, ...d_K)$ and the equation is given in equation (2.14):

$$b_m = \sum_{k=1}^{K} d_k a_m - 1 \tag{2.14}$$

The entropy is minimized entropy equations are given in equation (2.15):

$$d = \frac{\sum_{m=1}^{M} b_m^2}{\sum_{m=1}^{M} b_m^4} (A_0 A_0^R)^{-1} A_0 [b_1^3, b_2^3, \dots b_m^3] \tag{2.15}$$

where

$$A_0 = \begin{bmatrix} a_1 & a_2 & \cdots & a_M \\ 0 & a_1 & \cdots & a_{M-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{M-K+1} \end{bmatrix}_{K \times M} \tag{2.16}$$

where b_m is represented as the output of the impulse signal, and the original signal, d is represented as the arbitrary filter with length R . The equation (2.16) is known as the minimized spectral entropy, and it is employed to maximize the signal-to-noise ratio. The speech signal features are extracted. This method is used to overcome the redundant problem in the signals while identifying the speaker's voice, and these features are applied in real-time applications such as telephone banking and voice mailing. Then, the speakers' voices are classified by the Mask R-CNN classifier.

2.2.3 Classification of selected speaker features using Mask R-CNN classifier

Mask-R-CNN is defined as a simple, fully connected convolution NN architecture, which is used to classify the speaker's speech signals with high accuracy. It consists of three parts, such as ResNet, CNN, and a Fully connected layer with Mask.

2.2.3.1 ResNet-50/101-FPN

ResNet-50/101-FPN is the backbone of the Mask-RCNN architecture. The input speech signal is given to the ResNet layer. It is used to increase accuracy and speed. Then, the construction equation of the ResNet is given in equation (2.17):

$$b = D(a) + a \tag{2.17}$$

where a and b represents the input and output layers, $D(a)$ represents residual mapping functions and the equation is given in equation (2.18):

$$D(a) \cong V_2\tau(V_2 a + y_1) + y_2 \tag{2.18}$$

where τ is represented as the active function of the ReLU, $D(a)+a$ is the operation that is used as a short-cut connection with the element-wise addition. Its operation is to detect that the a and $D(a)$ are working as same. If it is not working in the same condition otherwise, choose V_s is selected as the short-cut connections for matching the dimensions and the equation is given as:

$$b = D(a) + V_s a \tag{2.19}$$

Equation (2.19) is known as the mapping function. Then, the feature mapping speech signal is given to the CNN network.

2.2.3.2 Mask-R-CNN

This method is employed to reduce the weights for a better classification process. The three layers that make it up are the convolutional, pooling, and fully connected ones. The fully connected layer consists of three branches, such as Mask, Softmax, and classification. The input signal is given as $A \in \aleph^{m \times n}$ is convoluted by the kernel sets L and it is given as $\{V_l \in \aleph^{w \times w}, L = 1,2,....L\}$ and the bias function is given as $\{y_l \in \aleph^{w \times w}, L = 1,2,....L\}$ are joined with the feature map function is given as A_L and generated with the element-wise nonlinear transform $\tau(.)$. This process is used for the construction of the CNN layer k, and the equation is given in equation (2.20):

$$A_l^k = \tau(V_l^k \otimes A^{k-1} + y_l^k) \tag{2.20}$$

where $\otimes$ is represented as the discrete convolutional operator. Here τ is represented as the weight parameter, and it is used to classify the speech signals on the basis of speaker input.

2.2.3.2.1 Pooling layer

The pooling layer is used to lower the samples or the down-sampling process. Then, the pooling layer equation is given in equation (2.21):

$$P_l^k = Pool(A_l^k) \tag{2.21}$$

here, the signals are trained by the process of convolution in an end-to-end supervised manner.

2.2.3.2.2 Sampled ROI

From the pooling layer feature map, process and trained as sampled ROI and the masking equation is given in equation (2.22):

$$K = K_{clocc} + K_{bbloss} + K_{mask} \tag{2.22}$$

The equation (2.22) is denoted as the multi-task loss function, where K_{clocc} and K_{bbloss} is represented as the classification loss and the bounding-box loss. Then the output of the ROI is connected with the fully connected layer with the mask branch Km^2. In this, τ parameter is used to minimize the binary cross entropy in the K_{mask} layer.

2.2.3.2.3 Fully connected layer

This is employed in the classification process. Here, the speech signals are classified in 3D arrays, such as $H \times W \times D$, H implies height, W implies width, D implies dimension. From

the softmax function, the speaker's ID is used to classify the speaker's speech signal and the softmax equation is given in equation (2.23):

$$soft(x) = \frac{expexp\,(x_l)}{\sum_{k=1}^{m} expexp\,(x_l)} \tag{2.23}$$

Then, the softmax layers of the speeches are in the ranges from (0, 1). Then, the speaker's ID are classified. It can increase the robustness of the speakers speech signal, and to get more accuracy, the weight parameters A, V, τ of Mask-R-CNN is optimized with the HCO algorithm.

2.2.4 Optimized the weight parameter of Mask R-CNN using Hosted Cuckoo Optimization (HCO) Algorithm

HCO is used to adjust the weight parameters of Mask-R-CNN to boost the accuracy and robustness of the voice signal. An effective optimisation technique is the HCO algorithm. The cuckoo's personality and way of life are outlined below. Laying eggs in the nest of another bird is in its nature. Because their eggs look and feel much like the host's, cuckoos will deposit their clutches on the host's nest. Many problems, including energy dispatch, system cost, usability, and cluster computing, can be improved with this optimisation algorithm. With the robustness of the speaker's voice signals, the HCO method is used in speaker identification and to improve the identification accuracy in this work. The Mask-R-CNN-HCO Algorithm's flowchart is shown in Figure 2.2. The stepwise procedure for the Hosted cuckoo optimization HCO algorithm is given below,

2.2.4.1 Hosted cuckoo optimisation (HCO) algorithm step-by-step approach

The weight parameters of Mask-R-CNN for speaker identification are optimized using the HCO technique. This enhancement increases the accuracy of identification by considering the robustness of the speaker's voice signals.

Step 1: Population initialization

Start by initializing the population of swarm nests $y_{j,l}$ each representing a solution for Mask-R-CNN with distinct parameter combinations, j is ranging from $(1\ to\ n)$, l is defined as the decision variable, population size denoted by n.

Step 2: Random Generation

Following the setup phase, the input parameters are randomly generated. The selection of the highest fitness values is contingent upon the weight parameter setting.

$$P(Y_{k,t}^{s}) = \frac{1}{1+e^{y_{k,t}^{s}}}, \qquad Y_{k,t}^{s} = \begin{cases} 0, & if\ P(Y_{k,t}^{s}) > \hat{v} \\ 1, & otherwise, \end{cases} \tag{2.24}$$

Step 3: Fitness Function

The fitness function values determine the solution of the random generation. The fitness function is used to optimise the weight parameter A,V,τ of Mask-R-CNN and achieve the goal function, as represented by the equation (2.25)

$$Fitnessfunction = optimizing(A,V,\tau) \tag{2.25}$$

Step 4: New solutions generation

The Levy fly technique is used to randomly pick the j^{th} solution, which can be stated as,

$$y_{j,l}(s+1) = y_{j.l}(s) + \alpha_1 \oplus Levy(\chi_1) \tag{2.26}$$

In the above equation, $y_{j,l}(s)$ represents the beginning position of the cuckoo, while the size parameter of the optimistic phase is given as α_1. Then, the random walk of levy distribution is denoted as χ_1

Step 5: Alien/foreign eggs detection

In order to locate the alien eggs, a probability matrix is created for every cuckoo solution. This is said or conveyed as,

$$\rho_{k,t} = \begin{cases} 0, & if\ (0,1) > \rho \\ 1 & otherwise \end{cases} \tag{2.27}$$

From the above equation, $\rho_{k,t}$ represents the detection of alien eggs. When the random number 0 and 1 is compared with ρ

Step 6: Remove worst habitats

A comparison is made between the fitness values $F_j\ and\ F_i$. With the optimum permutation of several parameters for the Mask-R-CNN, the highest value is acknowledged as the current finest solution. The lowest value is dropped from consideration, which can be represented as,

$$Step\_L = (perm\_random_2(Y_{k,t}^{s}) - perm\_random_1(Y_{k,t}^{s})) \times random\_number(0,1) \tag{2.28}$$

Step 7: Migration performance

In the end, the objective function is compared between the new solutions and the ones that are already in existence using the formula (2.29), and the solutions that are superior are selected for the subsequent generation

$$new\_solution = \begin{cases} Y_{k,t}^{s+1} & when\ R(Y_{k,t}^{s+1}) > R(Y_{k,t}^{s}) \\ Y_{k,t}^{s} & otherwise \end{cases} \tag{2.29}$$

From this equation, its migration performance increases that and the weight parameter A,V,τ of Mask-R-CNN can be optimized.

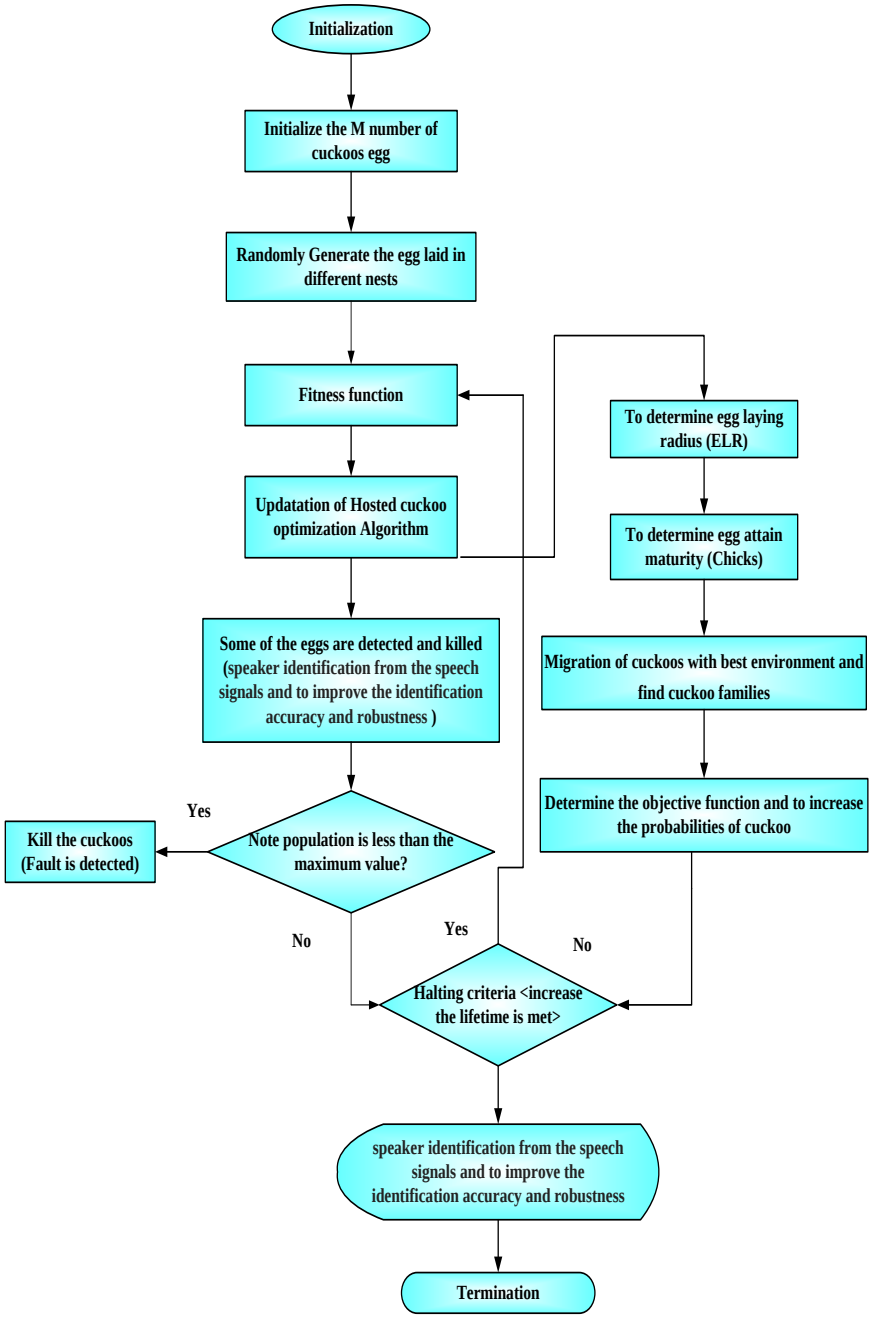

Figure 2.2 Mask-R-CNN-HCO Algorithm

Step 8: Termination

The optimal hyper-parameter has been chosen in Mask-R-CNN with the help of Improved HCO and can iteratively repeat step 3 until the halting criteria are met**.** Finally, the lifetime of the cuckoo is increased. In the same way, the robustness and the accuracy are increased and give speakers identity, and the Mask-R-CNN parameters are optimized using the HCO algorithm.

2.3 Results and Discussion

An effective framework for speaker identification is proposed, which would enhance the signal's accuracy and resilience via simulation. This framework would be built on Mask R-CNN classifier parameters adjusted using Hosted Cuckoo Optimisation (HCO). In order to identify the speaker, the suggested technique enhances both the identification precision and the speech signal's resilience. The recommended Mask-R-CNN-based model's performance and the different Test speakers' recall, accuracy, precision, sensitivity, and f-measure are the assessment metrics. Here, HCO is contrasted with five preexisting approaches, including Automatic Speaker Identification using K-Nearest Neighbours algorithm (KNN) [138], Multi-Class Support Vector Machine (MCSVM) [139], Gaussian Mixture Model - Convolutional Neural Network (GMMCNN) classifier [82], Deep Neural Network (DNN) [116], and Gaussian Mixture Model - deep Neural Network (GMMDNN) classifier [140]. The simulation parameters are presented in Table 2.1.

Table 2.1: ***Simulation Parameter***

Parameter	Value
MSR Identity Toolbox	1.0
GMM-based and I vector-based speaker identification systems	256 Gaussian components
Halting Criteria	0-1
Sampling Frequency	1600
Frame length	16 ms
Number of Speakers	120
Features	PNCC, Spectral Entropy, MFDPCC, GFCC

2.3.1 Evaluation Metrics

The various performance measures have been utilized to calculate the results for increasing the robustness of the signal. The specified terms are computed as,

2.3.1.1 False Acceptance Rate (FAR)

This is scaled as a rate, here illegal (Persons) accepted by the speech signals. Then, the FAR equation is given in equation (2.30):

$$FAR_{SI} = \frac{Number\,of\ \ persons\,accepted}{Total\,Number\,of\ \ person\,matching} \times 100\% \tag{2.30}$$

2.3.1.2 False Reject Rate (FRR)

This is computed as the probability of faults to identify the speech signal. Then, the FRR equation is given in equation (2.31):

$$FRR_{SI} = \frac{Number\,of\ \ signal\,\mathrm{Re}\ \ jected}{Total\,Number\,of\ \ image\,matching} \times 100\% \tag{2.31}$$

2.3.1.3 Error Rate (ER)

ER is defined as the FAR *i.e.* =FRR. Lowering the area gives better results. Then the ER equation is given in equation (2.32):

$$ER = (False\ Acceptance\ Rate\ (FAR) = False\ Rejection\ Rate\ (FRR) \tag{2.32}$$

2.3.1.4 Identification accuracy

This is computed as the maximal value of FAR along FRR. Then it is given in equation (2.33):

$$Accuracy = \frac{false\ \ accept\,rate + \ false\,reject\,rate}{2} \times 100 \tag{2.33}$$

2.3.2 Performance evaluation of various methods

Figures 2.3–2.8 show the performance metrics of the five current approaches—KNN, MCSVM, GMMCNN, DNN, and GMMDNN—as they compare to the Mask-R-CNN-HCO method's efficient speaker identification. Measures such as recall, specificity-score, accuracy, sensitivity, and precision are examined. By comparing the results using five different approaches, one can observe the effectiveness of the suggested system.

Figure 2.3 shows the performance metrics of the accuracy of the proposed Mask-R-CNN-HCO method. The proposed method is analyzed with five existing KNN, MCSVM, GMMCNN, DNN and GMMDNN methods. At speaker 10, the accuracy of the proposed method is 35.71%, 28.03%, 32.08%, 44.3%, and 24.35% superior to the existing KNN, MCSVM, GMMCNN, DNN and GMMDNN methods, respectively. At speaker 20, the accuracy of the proposed method is 34.71%, 27.03%, 35.08%, 42.3%, and 25.35% superior to the existing KNN, MCSVM, GMMCNN, DNN and GMMDNN methods, respectively. At speaker 30, the accuracy of the proposed method is 30.71%, 25.03%, 30.08%, 42.3%, and 34.35% superior to the existing KNN, MCSVM, GMMCNN, DNN and GMMDNN methods, respectively.

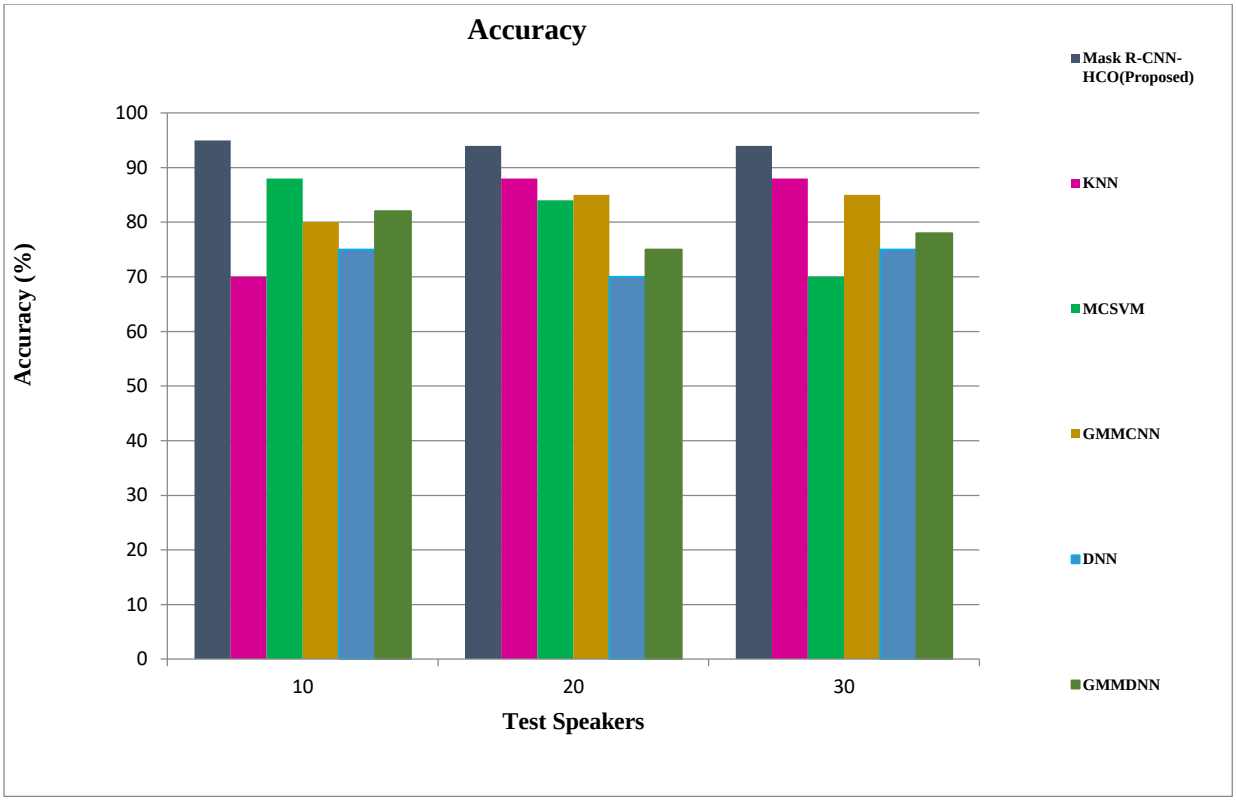

Figure 2.3 Performance metrics of accuracy

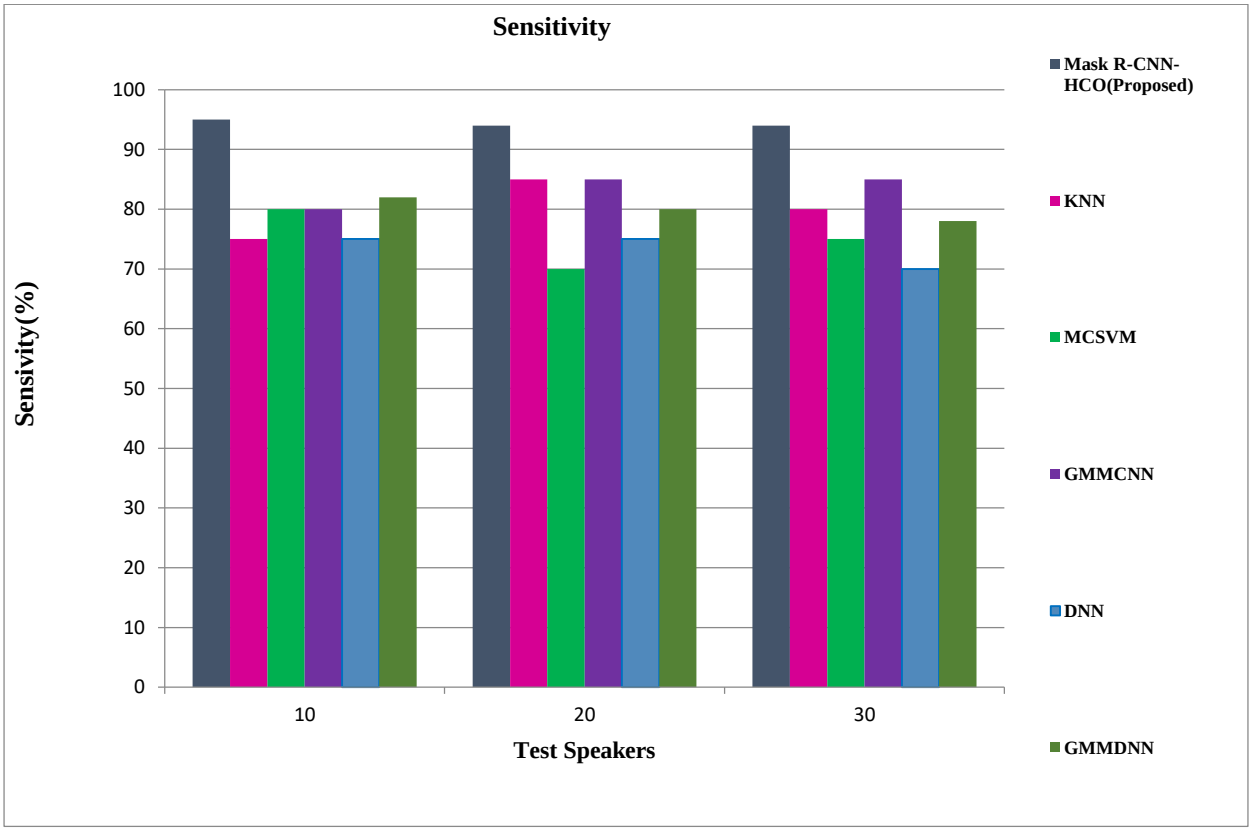

Figure 2.4 Performance metrics of sensitivity

Figure 2.4 shows the performance metrics of sensitivity of the proposed Mask-R-CNN-HCO method. The proposed method is analyzed with five existing KNN, MCSVM, GMMCNN, DNN, and GMMDNN methods. At speaker 10, the sensitivity of the suggested method is 32.71%, 38.03%, 22.08%, 42.3%, and 27.35% better than the existing KNN, MCSVM, GMMCNN, DNN, and GMMDNN methods, respectively. At speaker 20, the sensitivity of the proposed method is 24.71%, 29.03%, 38.08%, 45.3%, and 25.35% better than the existing KNN, MCSVM, GMMCNN, DNN and GMMDNN methods, respectively. At speaker 30, the sensitivity of the proposed method is 38.71%, 24.03%, 38.08%, 22.3%, and 36.35% better than the existing KNN, MCSVM, GMMCNN, DNN and GMMDNN methods, respectively.

Figure 2.5 shows the performance metrics of the precision of the suggested Mask-R-CNN-HCO method. The suggested method is compared with five existing methods, such as KNN, MCSVM, GMMCNN, DNN, and GMMDNN. At speaker ten, the Precision of the proposed method is 22.71%, 28.03%, 32.08%, 44.3%, and 25.35% higher than the existing method such as KNN, MCSVM, GMMCNN, DNN, and GMMDNN, respectively. At speaker 20, the Precision of the proposed method is 28.71%, 23.03%, 32.08%, 48.3%, and 26.35% higher than the existing method such as KNN, MCSVM, GMMCNN, DNN and GMMDNN, respectively. At speaker 30, the Precision of the proposed method is 37.71%, 28.03%, 33.08%, 21.3%, and 36.35% higher than the existing method such as KNN, MCSVM, GMMCNN, DNN and GMMDNN, respectively.

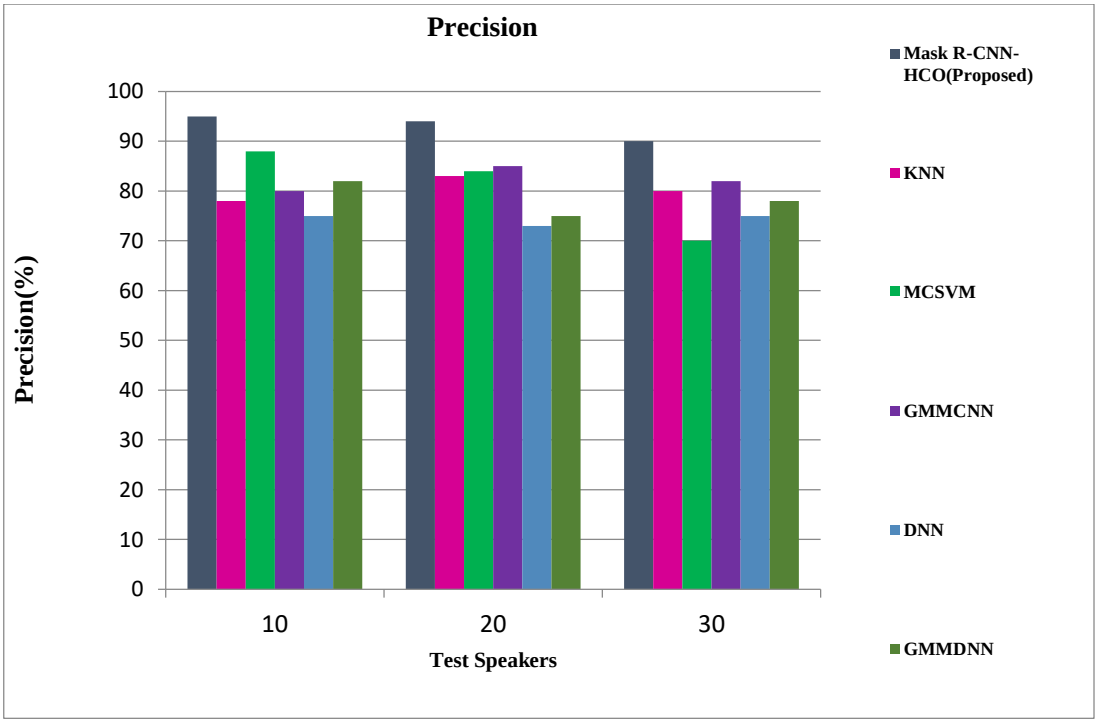

Figure 2.5 Performance metrics of precision

Figure 2.6 shows the performance metrics of the specificity of the proposed Mask-R-CNN-HCO method. The proposed method is analyzed with five existing KNN, MCSVM, GMMCNN, DNN and GMMDNN methods. At speaker ten, the specificity of the suggested method is 32.71%, 38.03%, 22.08%, 42.3%, and 27.35% higher than the existing method such as KNN, MCSVM, GMMCNN, DNN and GMMDNN, respectively. At speaker 20, the Specificity of the proposed method is 24.71%, 29.03%, 38.08%, 45.3%, and 25.35% higher than the existing method such as KNN, MCSVM, GMMCNN, DNN and GMMDNN, respectively. At speaker 30, the Specificity of the proposed method is 38.71%, 24.03%, 38.08%, 22.3%, and 36.35% higher than the existing method such as KNN, MCSVM, GMMCNN, DNN and GMMDNN, respectively.

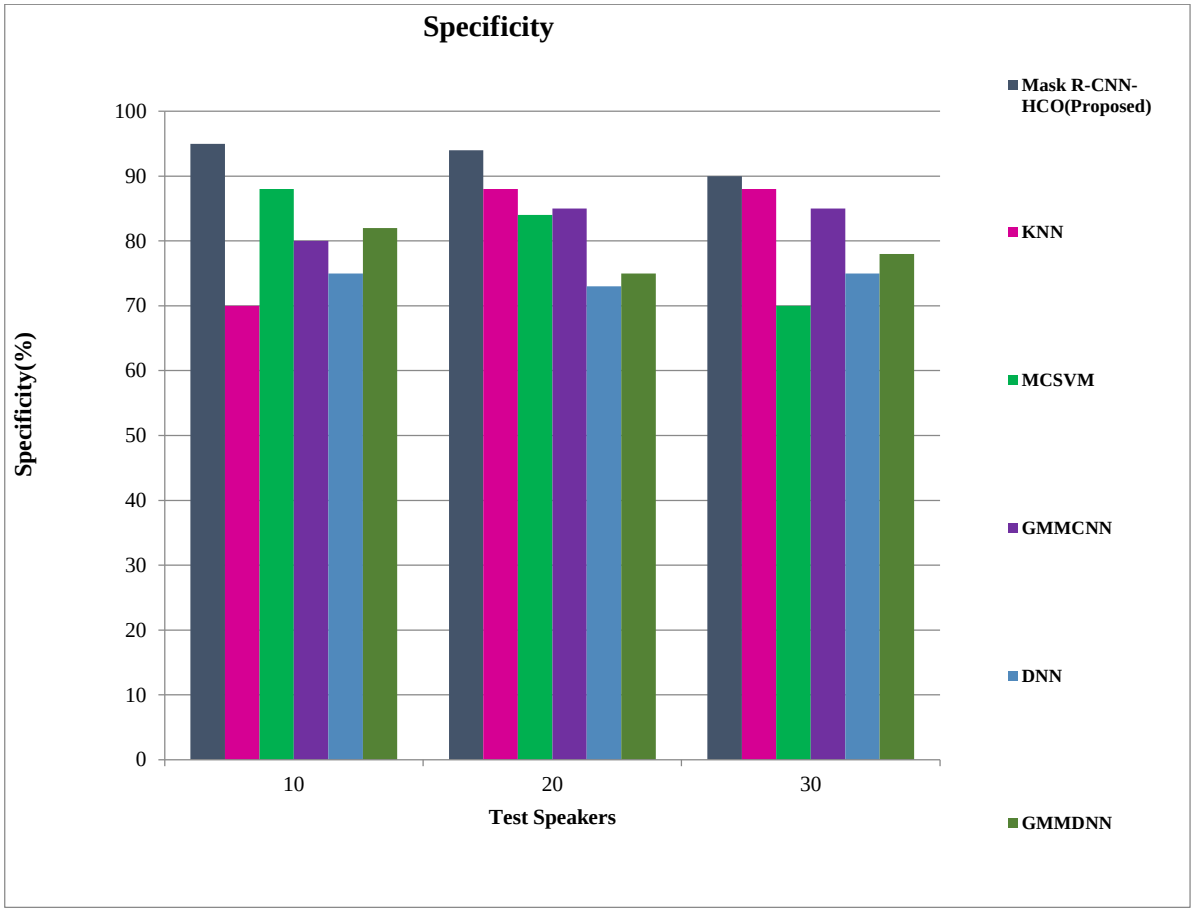

Figure 2.6 Performance metrics of specificity

Figure 2.7 shows the performance metrics of recall of the proposed Mask-R-CNN-HCO method. The proposed method is analyzed with five existing KNN, MCSVM, GMMCNN, DNN and GMMDNN methods. At speaker ten, the Recall of the proposed method is 32.71%, 38.03%, 22.08%, 42.3%, and 27.35% higher than the existing method such as KNN, MCSVM, GMMCNN, DNN and GMMDNN, respectively. At speaker 20, the Recall of the proposed

method is 24.71%, 29.03%, 38.08%, 45.3%, and 25.35% higher than the existing methods, such as KNN, MCSVM, GMMCNN, DNN and GMMDNN, respectively. At speaker 30, the Recall of the proposed method is 38.71%, 24.03%, 38.08%, 22.3%, and 36.35% higher than the existing methods, such as KNN, MCSVM, GMMCNN, DNN and GMMDNN, respectively.

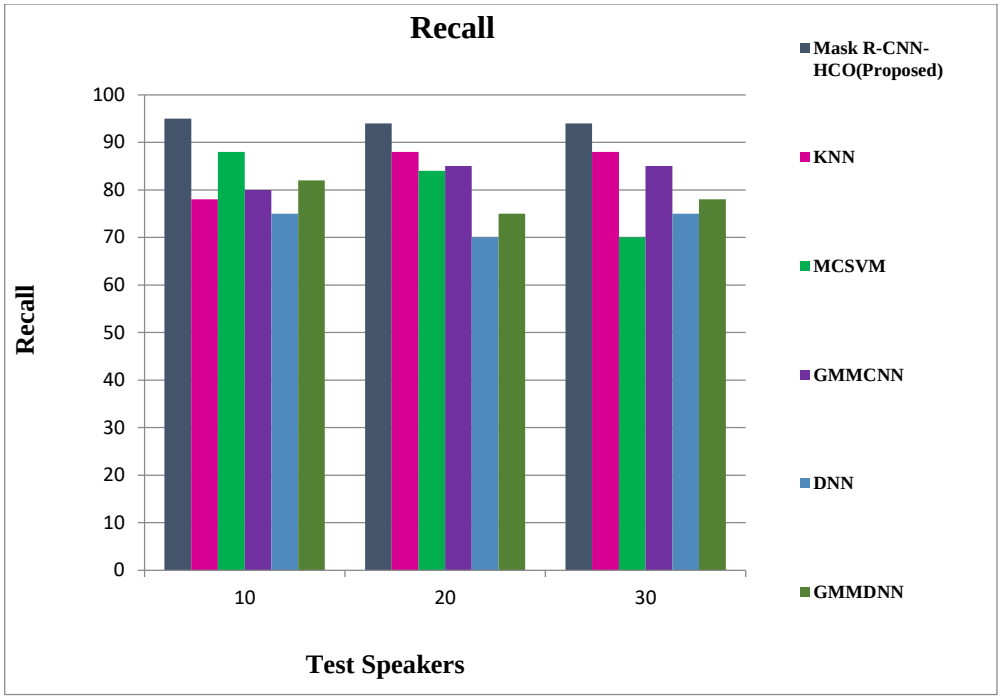

Figure 2.7 Performance metrics of recall

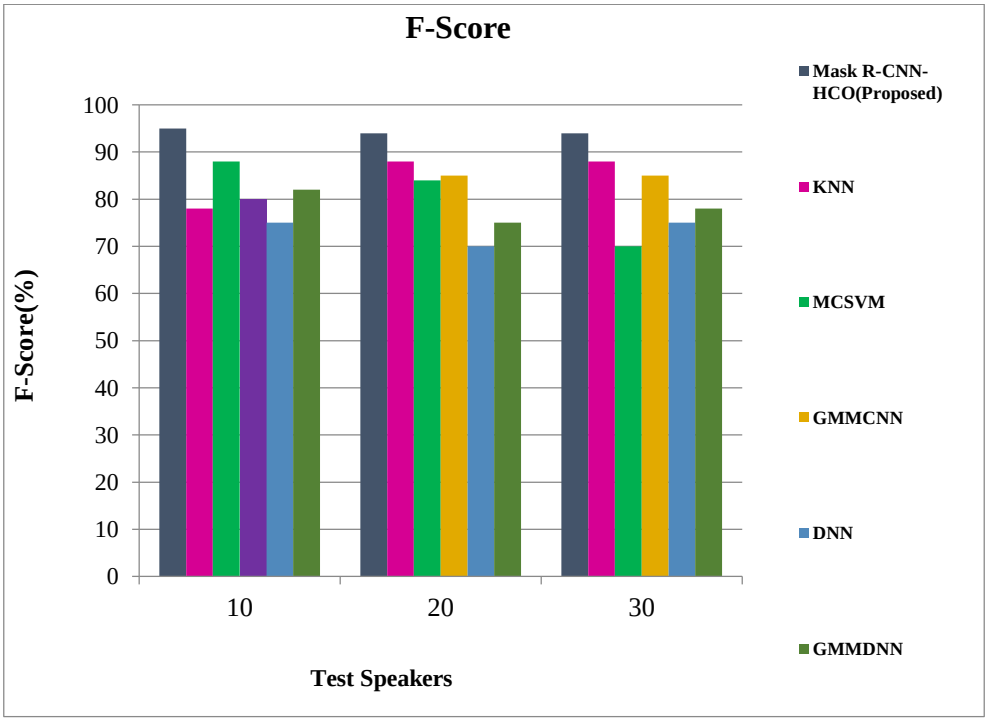

Figure 2.8 Performance metrics of F-score

Figure 2.8 shows the performance metrics of the F-score of the proposed Mask-R-CNN-HCO method. The proposed method is analyzed with five existing KNN, MCSVM, GMMCNN, DNN and GMMDNN methods. At speaker 10, the F-score of the proposed method is 32.71%, 38.03%, 22.08%, 42.3%, and 27.35% superior to the existing KNN, MCSVM, GMMCNN, DNN and GMMDNN methods, respectively. At speaker 20, the F-score of Mask-R-CNN-HCO is 24.71%, 29.03%, 38.08%, 45.3%, and 25.35% superior to the existing KNN, MCSVM, GMMCNN, DNN and GMMDNN methods respectively. At speaker 30, the F-score of Mask-R-CNN-HCO is 38.71%, 24.03%, 38.08%, 22.3%, and 36.35% superior to the existing KNN, MCSVM, GMMCNN, DNN and GMMDNN methods respectively.

A confusion matrix in the context of Mask-R-CNN-HCO (Figure 2.9) illustrates the performance of the model in terms of classification outcomes. This matrix is particularly useful for evaluating the accuracy of the model's predictions.

The data's actual classes or categories are shown in the rows of the confusion matrix, while the Mask-R-CNN-HCO model's predicted classes are shown in the columns. Instances where the model's predictions align with the actual classes, indicating accurate classifications, are represented by the major diagonal of the matrix. Off-diagonal elements represent misclassifications, where the predicted and true classes differ. The confusion matrix provides a comprehensive view of the model's performance by breaking down the results into four categories:

True Positives (TP): Instances where the model correctly predicts positive classes.

True Negatives (TN): Instances where the model correctly predicts negative classes.

False Positives (FP): Instances where the model incorrectly predicts positive classes.

False Negatives (FN): Instances where the model incorrectly predicts negative classes.

From these values, various performance metrics such as accuracy, precision, recall, and F1 score are calculated, offering insights into the overall effectiveness of the Mask-R-CNN-HCO model.

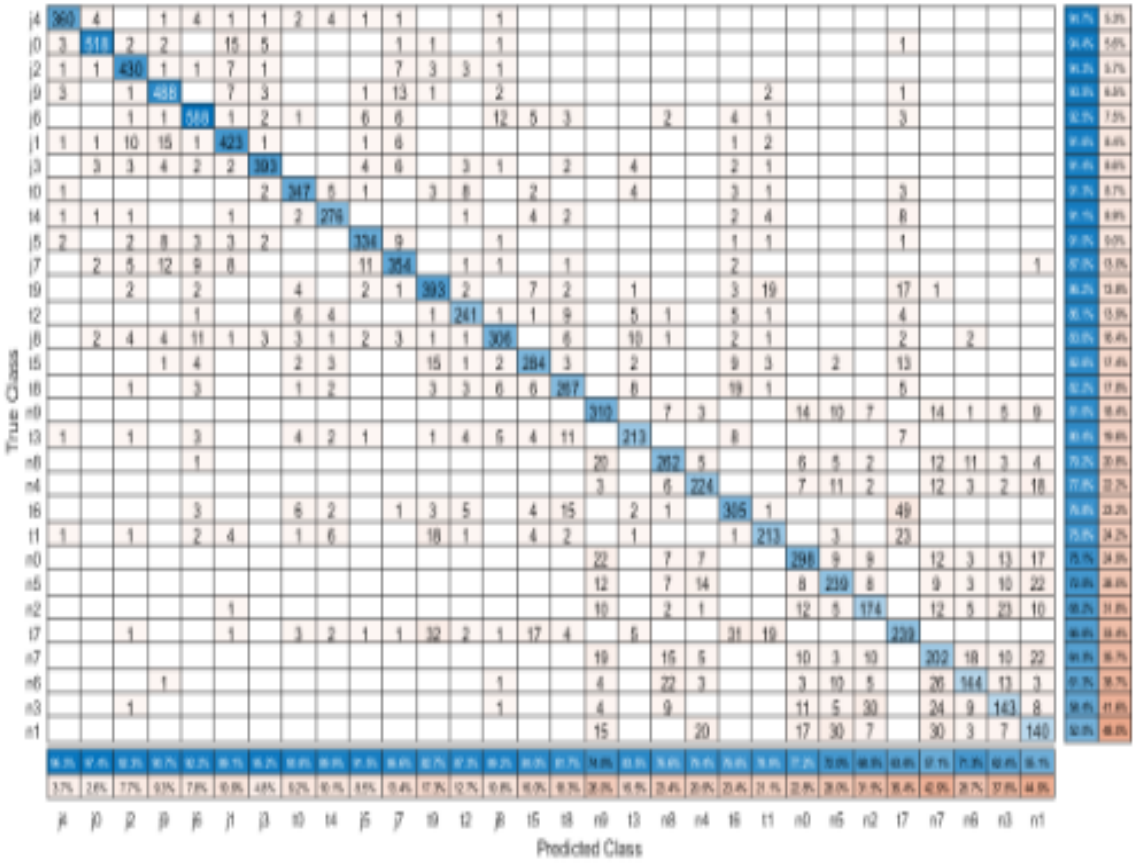

Figure 2.9 Confusion matrix of proposed Mask-R-CNN-HCO

2.4 Real-time applications

Here, speech signals are identified from different test speakers consisting of speech signals such as telephone speech from telephone banking (TB) and the speech from internet videos from fax mailing (FM). The performances real-time applications are given in equations (2.34 and 2.35):

$$RL_{TB\ and\ FM} = 0.25\left(RL_n\left(RL_{TB}, K^{(1)}_{target\ factor}\right) + RL_n\left(RL_{TB}, K^{(2)}_{target\ factor}\right) + RL_n\left(RL_{TB}, K^{(3)}_{target\ factor}\right)\right) \quad (2.34)$$

where,

$$RL_n = K_{submission} + \left(\frac{RL_{submission}}{RL_{scaling\ factor}} . \frac{1-RL_{target\ factor}}{RL_{target\ factor}}\right) RL_{scaling\ factor} \quad (2.35)$$

where RL is represented as real-time applications, TB is represented as telephone banking, $_{TB}$ is represented as fax mailing. $RL_{submission}$ and $RL_{traget\ factor}$ are the application parameter, and it is set as 1 and the probabilities of the targets are set as $K^{(1)}_{traget\ factor} = 0.01$, $K^{(2)}_{traget\ factor} = 0.005$,

$K^{(3)}_{traget\ factor} = 0.05$ and $RL_{traget\ factor}$, RL_{TB} is represented as detection cost with the target factor $K^{(1)}_{traget\ factor}$ and $K^{(2)}_{traget\ factor}$ and it is applicable in telephone banking, and the next factor is $0.5RL_{FM}$ calculated the half of the primary cost, and it is applicable in the fax mailing. In this manner, this Mask-R-CNN-HCO identifies the speakers of the telephone banking and fax mailing and is utilized as a real-time application.

2.5 Chapter Summary

An effective method for speaker identification is suggested, which enhances accuracy and signal resilience by employing parameters adjusted via Hosted Cuckoo Optimisation (HCO) for Mask R-CNN classifiers. Prior to processing, the incoming voice signals are checked against the database of real-time speakers. We take the input speech signal and use four features—MFDPCC, GFCC, PNCC, and spectral entropy—to make the signal more robust. A speaker's ID is then classified using the Mask R-CNN classifier. We next optimise the parameters of the Mask R-CNN classifier using the HCO method. Applications that operate in real-time, including faxing and telephone banking, subsequently use the approach. In MATLAB, the simulation is executed. The proposed Mask-R-CNN-HCO method attains Recall of 22.43%, 32.41%,39.13%, 33.80%, 32.43%, Specificity 12.81%,22.43%, 32.41%,39.13%, 7.60%, F-Score 22.43%, 32.41%,39.13%, 24.3%, 35.71% higher than the existing methods such as Automatic Classification of speaker identification using KNN, MCSVM, GMMCNN, DNN, GMMDNN.

3. Chapter - Speaker Verification

One area where using Speaker Verification might be hard is in forensics. This usually happens after a crime has already taken place, when it becomes critical to confirm the criminal's identity by matching their voice to a recorded message. Traditionally, this task was accomplished by educating an expert who had the ability to discern the speaker's voice by analysing the visual characteristics of their speech, such as spectrograms and voice prints. However, the reliability and effectiveness of these strategies were found to be inadequate. In order to establish the suspect's guilt, it is necessary to conclusively confirm that the vocal characteristics of the criminal and the suspect are identical, leaving no room for reasonable doubt. In order to address this issue, there is a need for an automated and dependable Speaker Verification system. Speaker Verification (SV) is the procedure of verifying the identification of a speaker by analysing their voice data [65] [111]. Figure 3.1 presents a generic block diagram of a Speaker Verification System (SVS) [142]. When a speaker declares their identification, their voice sample is compared to the model linked to the asserted identity.

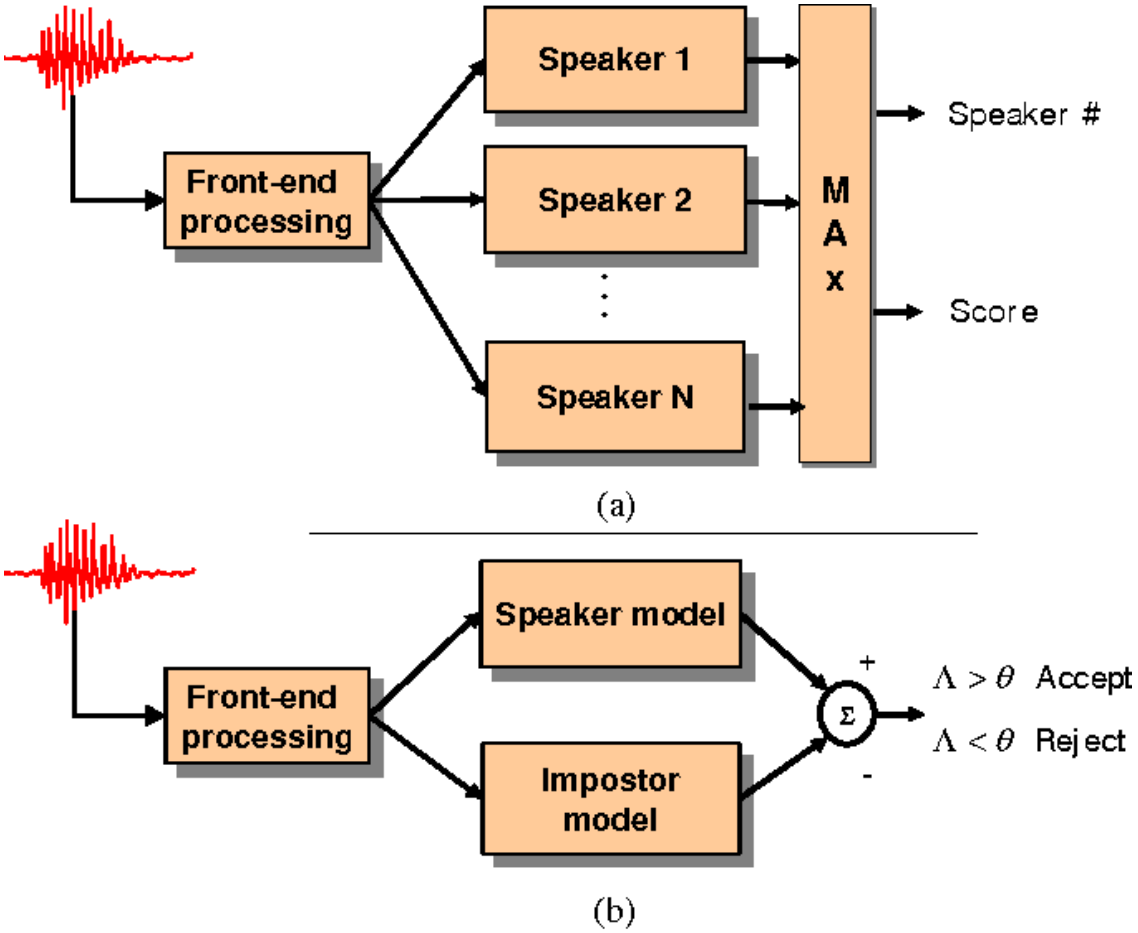

Figure 3.1 Block diagram (a) Speaker Identification (b) Speaker verification [142]

A threshold notion is used by the system to make a decision. When the difference between the test speech data and the claimed identity speaker's model drops below a certain threshold, it indicates that the speaker may be believed to be authentic. Essentially, the system assesses the degree of similarity between the attributes of the given speech and the traits associated with the claimed identity. If the level of similarity is sufficiently high (*i.e.*, below the specified limit), the system acknowledges the speaker as authentic.

It is important to emphasise that Speaker Verification and Speaker Identification are distinct from each other (figure 3.1). Speaker Verification yields a binary outcome: either accepting or rejecting the asserted identification. This decision is based on whether the distance between the target model and the test speech data is less than a certain threshold. Notably, a Speaker Verification System's performance is independent of the total number of speakers in the dataset. This method is mainly concerned with confirming or refuting the given identification and pays little attention to the size of the overall collection of speakers. The decision-making process is made simpler and its efficacy is guaranteed to be unaffected by the number of speakers in the dataset thanks to its binary structure.

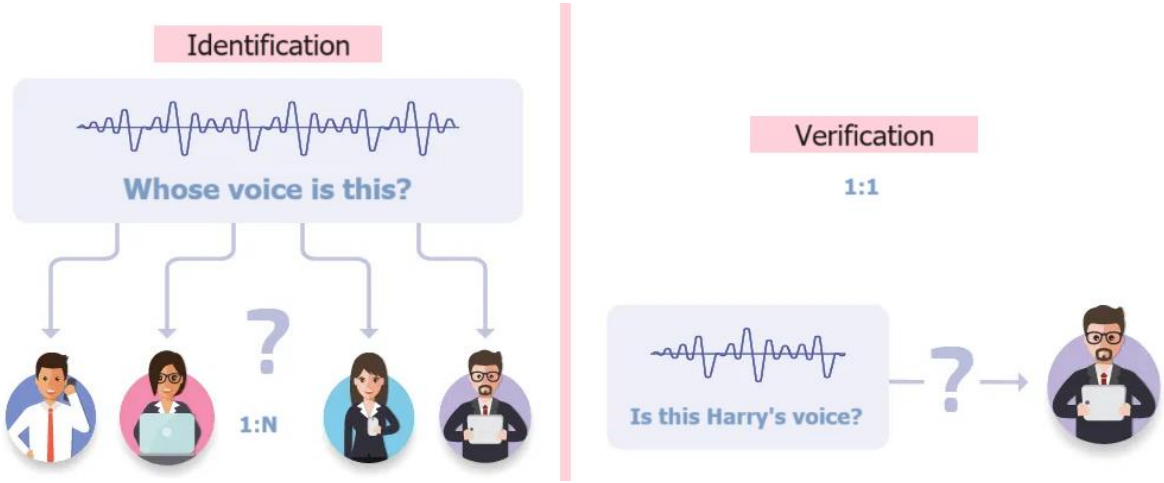

Figure 3.2 ***Speaker Identification and Speaker Verification***

Speaker verification has seen significant improvement over time. Here is a concise summary of its progression:

Early Years (1960s-1990s): The origins of speaker verification may be attributed to the 1960s, during which academics initiated investigations into the distinctive characteristics of the human voice. Initial systems depended on basic attributes such as pitch and duration. Nevertheless, their progress was hindered by the absence of advanced methods and computational capabilities. Signal processing methods had notable breakthroughs throughout the 1990s, enabling researchers to extract more intricate information from speech transmissions. Mel-

frequency cepstral coefficients (MFCCs) have gained popularity as a way for accurately expressing the spectrum properties of speech, hence enhancing the accuracy of speaker verification.

The emergence of Statistical Models in the 2000s: In the 2000s, statistical models such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) were used for speaker verification. These models provide a stronger and more comprehensive structure for capturing the diversity in speech signals and differentiating across speakers. The emergence of machine learning in the 2010s led to the use of increasingly sophisticated approaches, including Support Vector Machines (SVMs), deep neural networks (DNNs), and convolutional neural networks (CNNs), in speaker verification systems. These methods enhanced the ability of speaker verification systems to accurately distinguish between speakers, particularly when dealing with bigger datasets.

Deep Learning Dominance (Late 2010s-2020s): Deep learning, particularly deep neural networks (DNNs) and neural embeddings, has emerged as the prevailing approach in speech verification in recent years. Speaker embeddings, which are trained on extensive datasets, have shown exceptional performance, allowing for speaker verification in practical scenarios.

However, speaker verification encounters several challenges:

Data Variability: Speaker verification systems may have difficulty handling the variability in voice signals due to variables such as ambient noise, emotional state, or physiological parameters. It is essential to adjust to these fluctuations in order to implement them in real-life situations successfully.

Data security and privacy concerns: These issues have escalated due to the extensive use of biometric technology. Safeguarding voice data and guaranteeing the continuous implementation of secure enrollment and verification procedures are persistent obstacles.

Spoofing and impersonation: Adversaries may use spoofing and impersonation techniques to trick speaker verification systems by using pre-recorded or artificially generated speech samples. Creating effective techniques to avoid spoofing assaults is an ongoing and difficult task.

Non-standardization: The lack of standardised criteria and datasets poses difficulties in systematically comparing and evaluating various speaker verification systems.

Ethical and Legal Considerations: The use of biometric technology, such as speaker verification, gives rise to ethical and legal concerns pertaining to permission, monitoring, and

the possibility of unauthorised exploitation. Achieving a harmonious equilibrium between the need for security and the protection of individual rights is a continuous and persistent problem. Speaker verification has progressed from basic signal processing methods to advanced deep learning models. Nevertheless, the field's progress is still influenced by obstacles such as data unpredictability, security concerns, spoofing, standardisation, and ethical issues.

3.1 Speaker verification model framework

Speaker verification faces a crucial obstacle in identifying the most effective method of extracting features, as well as selecting a suitable classifier to differentiate between authentic and falsified speech samples. This differentiation is vital for applications such as forensic analysis. Diverse approaches have been investigated to tackle such problems and improve the overall efficiency of speaker verification systems. The use of constant-Q cepstral coefficients is one such example; these coefficients provide weight to the magnitude characteristics of the speech signal in order to prioritise the extraction of unique features. We also integrate modulation dynamic features with Mel-scale and gamma-tone scale filters, and we employ glottal closure instants to capture temporal dynamics. In order to improve the system's overall performance, these characteristics are combined with sets like Constant-Q Cepstral Coefficients (CQCC) and Mel-Frequency Cepstral Coefficients (MFCC).

Various methods of combining scores have been used to improve the correctness of the system. An example of this is the utilisation of the variable-length Energy Separation Algorithm (VESA) in conjunction with Identical-Frequency Cepstral Coefficients (IFCC) and different fusion systems. These fusion systems include epoch features with Gaussian Mixture Model (GMM), IFCC with GMM, MFCC with GMM, and CQCC with GMM.

The dissertation presents a self-contained approach to further enhance the system's performance. This entails the introduction of an innovative technique or enhancement to the current techniques used for feature extraction, classification, or fusion. The aim is to make contributions to the progress of speaker verification systems, particularly in the field of forensic applications. A two-tier feature extraction framework in a speaker verification model involves a multi-stage process to extract relevant features from speech data. This method shines in fields like forensics, where confirming the speakers' identities is of the utmost importance for investigations.

The proposed TTFEM-AFSV model, which stands for Two-Tier Feature Extraction with Metaheuristics Automated Forensic Speaker Verification, employs the average median

filtering (AMF) method to eliminate background noise from audio data. The Inception v3 model, which is built on a deep convolutional neural network, takes MFCC and spectrograms as initial inputs. In addition, the Inception v3 model's hyperparameters are fine-tuned using the ant lion optimizer (ALO) algorithms. The last step in automated ASR is to use a classifier that takes use of a long short-term memory with a recurrent neural network (LSTM-RNN) mechanism. The validity of the TTFEM-AFSV model's performance is examined by a battery of tests.

3.1.1 State of art

A person's voice may be used as a biometric in speaker recognition. Because the larynx, vocal tract, and other parts of the voice producing organ vary in size and structure, each sound would be unique [143]. The fact that the test speaker is not revealing their identify adds another layer of complexity to this method. The procedure needs to sort the speakers in a 1:N ratio, where N is the total number of speakers. Assuring that a speaker is who they say they are is known as speaker verification. It is referred to as open-set recognition since it is believed that a false claim cannot be discovered [144]. Generally, it is regarded as the test speaker's voice coming from the known speaker; also, it is called closed-set recognition. A human speech signal is a stronger medium of transmission that contains rich information, namely emotional traits, accent, gender, and so on. These unique features enable researchers to recognize speakers through voiceprint identification [145]. The collected speaker utterance is fed into the deep learning network for training. In the identification technique, the speaker recognition method matches the extracted speaker features with those in the model library [146]. Next, the target speaker is determined by the speaker's likelihood of utterance.

An individual's speech signal may be definitely identified using the automated speaker recognition method. Law enforcement monitoring, specifically security control web service and private data area, forensics for crime investigation, remote time logging, and prison call observation are some of the many uses for automatic speaker recognition method. Other examples include controlling access to services like voice mail and telephone banking, data networks voice dialling, database access services, remote access to computers, information services, and telephone shopping [147]. It is essential to extract the feature from every frame that captures key features of the speech signal for the speaker verification system [148]. Linear prediction cepstral coefficient (LPCC), perceptual linear predictive coefficient (PLPC), and Mel frequency cepstral coefficients (MFCC) are different feature extraction techniques utilized in speaker verification systems. The MFCC is extensively employed as the feature extraction

technique for the current speaker verification system, and it accomplished higher efficiency in clean conditions [149]. However, the efficiency of the MFCC feature drops considerably in reverberation and noise situations [150]. The major problem with the present techniques is the size of every input sample is massive enough to identify the speaker and correctly train, thus improving the requirement of the resources [151]. Also, it failed to work with optimum performance in noisy environments. The study developed a strong speaker identification technique using a hybrid mechanism for automatically recognizing the speaker from speech signals efficiently [152].

3.2 TTFEM-AFSV model

A TTFEM-AFSV model has been developed to verify speakers in forensic applications. The TTFEM-AFSV model initially exploited the AMF technique to discard the noise present in speech signals. Next, MFCC and spectrograms are considered inputs to the Inception v3 model as metaheuristic approach [153]. Then, the ALO algorithm is utilized to fine-tune the hyperparameters related to the Inception v3 model. At last, the LSTM-RNN model is exploited as a classifier for automated ASR. Figure 3.3 depicts the overall process of the TTFEM-AFSV approach.

3.2.1 Pre-processing

The median filter (MF) retains the sharpness of an image and eliminates the noise. All the pixels are substituted with median values from the neighbouring pixel. This filter utilizes a 3 × 3 window [154]. It is the best filter among traditional filters that eliminate the speckle noises. The step followed to create the MF is shown in Algorithm 1. Spatial processing to retain the edge details and to remove non-impulsive noises through the adaptive MF plays a crucial part. The AMF preserves the smaller structure in the image and edge. In the AMF, the window size differs with regard to every pixel.

3.2.2 MFCC and Spectrograms

Spectrograms and MFCC are the two speech features utilized in this work. MFCC depends on a cepstral analysis of speech signals and is extensively applied in different speech applications [155]. The segmentation of speech samples into 25 ms frames with a 10-ms frameshift between adjacent occurrences is the basis of the cepstral analysis. The Fast Fourier Transform (FFT) is applied after the frames are multiplied by the Hamming window. A triangular Mel-filter bank is used to filter the FFT response. It follows that MFCC is applied to the filter bank response in order to get Discrete Cosine Transform characteristics. Based on frequency domain analysis

of the voice input, MFCC is normalised with mean and variance to generate a spectrogram with unit variance and zero mean. A hamming window with a step size of 10 m and a length of 25 ms is used for the calculation in a sliding window approach. Spectogram features with zero mean and unit variance are obtained by applying normalised mean and variance to the data.

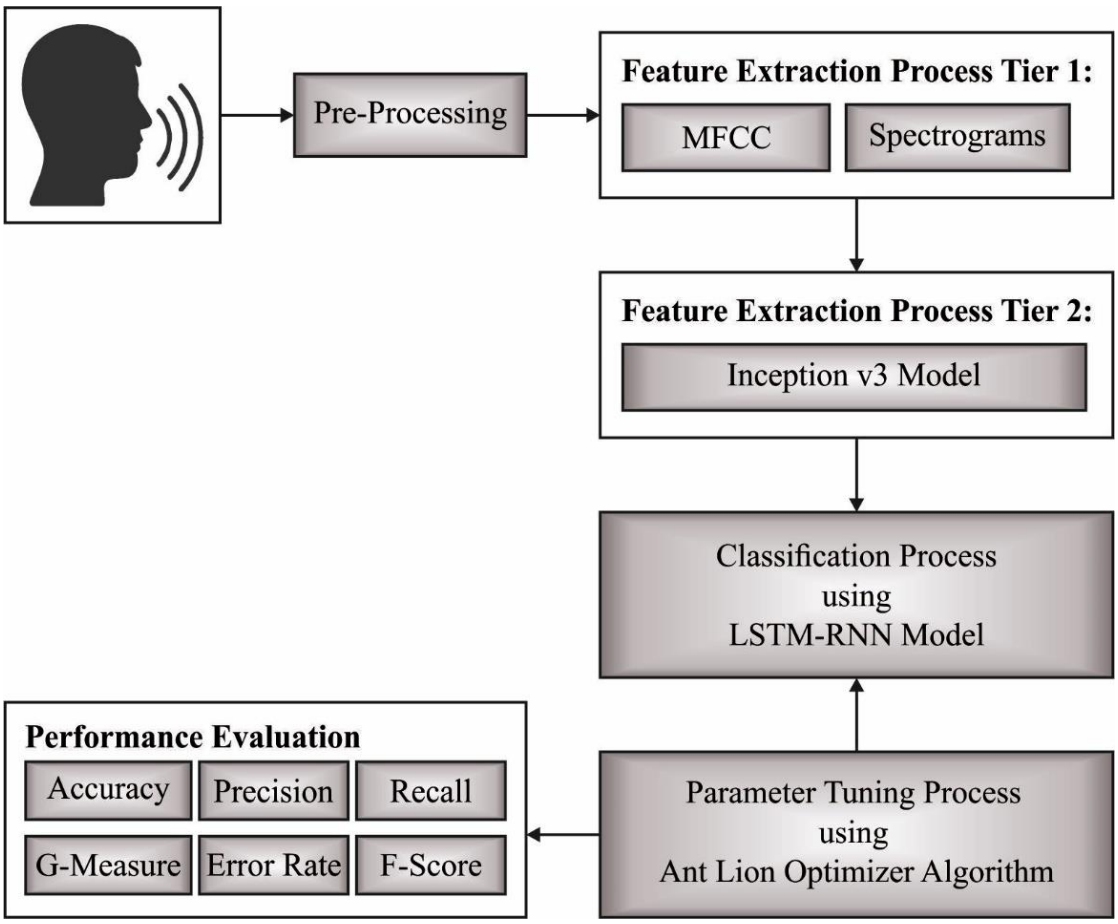

Figure 3.3 The TTFEM-AFSV model

Algorithm 1: Pseudocode of AMF
(1) Consider "A" input matrix with N columns and M rows. (2) Create a matrix with N + 2 columns and M + 2 rows by adding a zero to the side of the input matrix (3) Take a mask of size 3 × 3. (4) Place the mask on the initial component, viz., the first column and row of matrix "A." (5) Select each element listed with the mask and arrange them in ascending sequence. (6) Take the median value (center component) from the sorted array and substitute the component A(1, 1) with the median values (7) Slide the mask to the following component.

(8) Reiterate the 4 to 7 steps until each matrix "A" element is substituted with the respective median values.

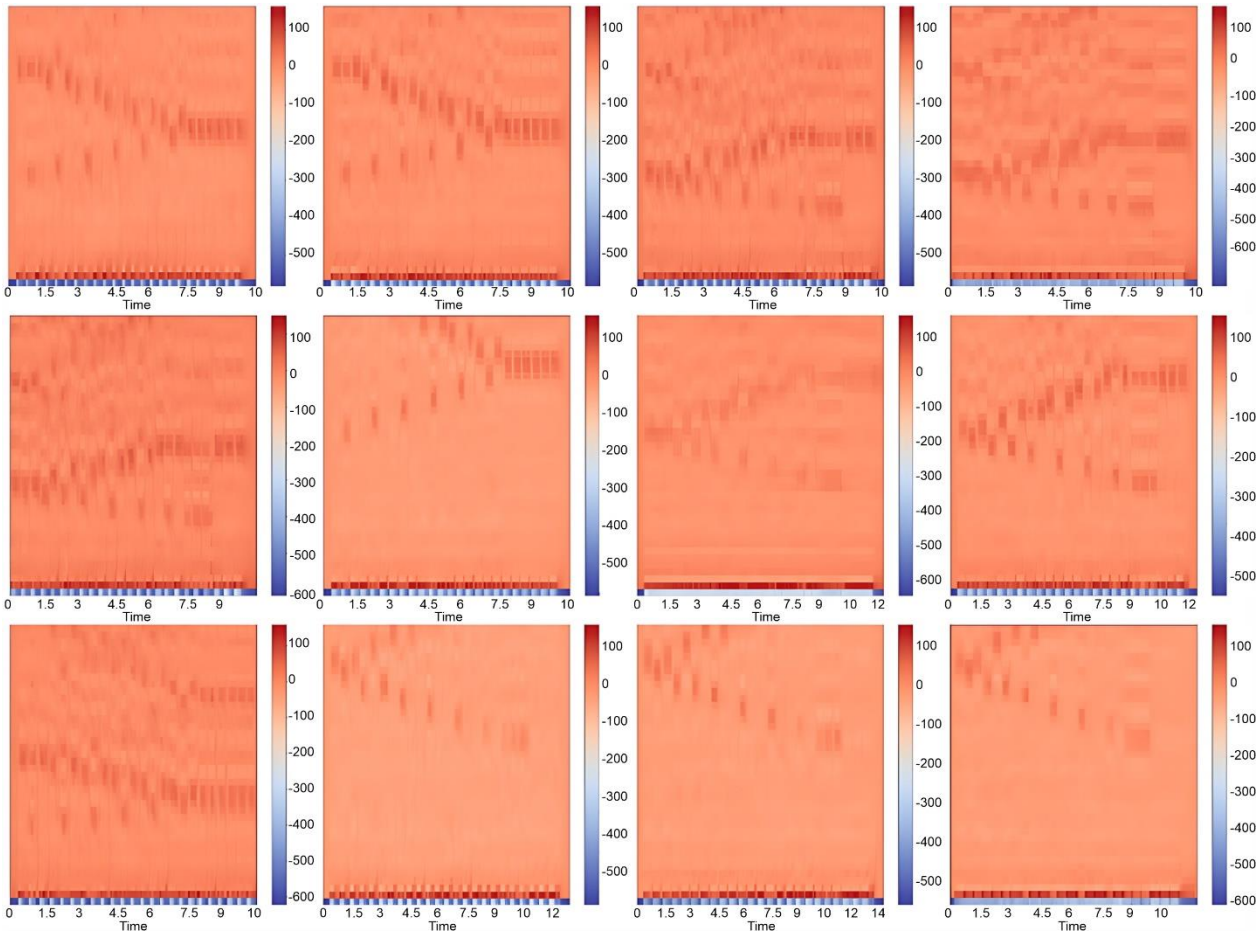

Figure 3.4 Sample extracted features

3.2.3 Inception v3-based feature extraction

The presented method uses the MFCC spectrogram as CNN's input; the spectrogram is a two-dimensional signal and comprises the identity data of the speaker. Simultaneously, CNN provides translation invariance in time and space. Thus, the voiceprint feature could be obtained within the spectrogram space without interrupting the time series. As a result, this study presents the MFCC and spectrogram as the input of the Inception_v3 module [156]. Inception-v3 is an alternative to the presented GoogLeNet image recognition Inception structure. The Inception modules generally contain dissimilar sizes of convolution and maximal grouping layers. Add a channel to the network output of the preceding layer over convolutions and later implement a non-linear combination. Overfitting might be evaded, and the network adaptability and expression to various scales are enhanced. Inception_v3 is a Keras-developed network architecture that is trained in ImageNet. Figure 3.4 illustrates the

sample feature extraction. The image input is 299 × 299-pixel size with three channels. In contrast to the preceding version (Inception v1 and v2), the Inception_v3 structure uses the convolutional kernel partitioning model for dividing larger integrals into small convolutions. For instance, a 3 × 3 line is separated into 1 x 3 and 3 x 1 lines. The segmentation model is utilized for minimizing the parameter count. While increasing the speed of the training network, it efficiently retrieves spatial features. Simultaneously, Inception v3 improves the Inception structure with 8 × 8, 17 × 17, and 35 × 35 regional grid sizes.

3.2.4 ALO-based hyperparameter tuning

In this study, the ALO algorithm is used to adjust the hyperparameters of the Inception v3 model. In this study, the ALO procedure has been applied [157]. The lifecycle of an ant-lion comprises two major phases: adulthood and larva. Ant-lion larva digs a conically formed hole using their hand by stirring in a round sphere and discard the soil using the jaw. In this work, the activities of ants as prey and ant-lion as hunters are modelled. For modelling such encounters, the ant should move near the searching region, and the ant-lion hunt them. Ant-lion fitness increases when they set up further traps. Meanwhile, the ant moves randomly to find food; an arbitrary move to model and motion is defined as follows

$$x(t) = [0, totalsum(2r(t_1) - 1), \dots, totalsum(2r(t_n) - 1)] \tag{3.1}$$

$$r(t) = \begin{cases} 1 & if\ rand > 0.5 \\ 0 & if\ rand \leq 0.5 \end{cases} \tag{3.2}$$

From the equation, the overall sum illustrates the overall density, n represents the maximal amount of iterations, t denotes the random motion phase. $r(t)$ shows the random function, and $rand$ indicates an arbitrary value produced by uniform distribution within zero and one. The position of the ant is saved and is later applied in the optimization:

$$M_{Ant} = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,d} \\ A_{2,1} & A_{2,2} & \dots & A_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ A_{n,1} & A_{n,2} & \dots & A_{n,d} \end{bmatrix} \tag{3.3}$$

$$M_{OA} = \begin{bmatrix} f([A_{1,1} & A_{1,2} & \dots & A_{1,d}]) \\ f([A_{2,1} & A_{2,2} & \dots & A_{2,d}]) \\ \vdots & \vdots & \vdots & \vdots \\ f([A_{n,1} & A_{n,2} & \dots & A_{n,d}]) \end{bmatrix} \tag{3.4}$$

From the equation, M_{Ant} Denotes the matrix that comprises the position of every ant, $A_{i,j}$ Indicates the value of $j\text{-}th$ variables, n specifies the ant count, d indicates the variable count. Storing the position and the fitness of the ant-lion are shown as follows:

$$M_{Ant-lion} = \begin{bmatrix} ALi_{1,1} & ALi_{1,2} & \dots & ALi_{1,d} \\ ALi_{2,1} & ALi_{2,2} & \dots & ALi_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ ALi_{n,1} & ALi_{n,2} & \dots & ALi_{n,d} \end{bmatrix} \quad (3.5)$$

$$M_{OAL} = \begin{bmatrix} f([ALi_{1,1} & ALi_{1,2} & \dots & ALi_{1,d}]) \\ f([ALi_{2,1} & ALi_{2,2} & \dots & ALi_{2,d}]) \\ \vdots & \vdots & \vdots & \vdots \\ f([ALi_{n,1} & ALi_{n,2} & \dots & ALi_{n,d}]) \end{bmatrix} \quad (3.6)$$

Now, $M_{Ant-lion}$ denotes the matrix encompassing the position of every ant-lion, $ALi_{i,j}$ Indicates the value of $j\text{-}th$ parameter, n characterizes the ant-lion count, and d displays the variable count. M_{OAL} Denotes the matrix encompassing the fitness of every ant-lion, and f demonstrates the targeted function.

The ant upgrades the position by arbitrarily moving at every optimization stage. Since each searching space has a border, a normalizing process is employed for the random movement to keep them within the interval of the searching region. Here, the $Min\text{-}Max$ normalization technique is applied:

$$X_i^t = \frac{(X_i^t - a_i) * (d_i^t - c_i^t)}{(b_i^t - a_i)} + c_i \quad (3.7)$$

In Eq. (3.7), b_i and a_i Shows the maximal and the minimal arbitrary motion of the $j\text{-}th$ parameter, correspondingly. Furthermore, d_i^t andc_i^t Indicates the maximal and minimal of the $i\text{-}th$ parameter in the $t\text{-}th$ iterations, correspondingly. The formula should be iteratively applied to guarantee that the arbitrary movement is reserved.

Ant-lion traps affect the arbitrary motion of ants. For mathematically modeling the assumption, the subsequent equation is applied:

$$c_i^t = Ant - lion_j^t + c^t \; d_i^t = Ant - lion_j^t + d^t \quad (3.8)$$

In Eq. (3.8), d^t and c^t Shows the vector comprising the maximal and minimal value of each variable in the $t\text{-}th$ iterations, correspondingly. Furthermore, d_i^t and c_i^t Denotes the maximal and the minimal value of each variable of the $j\text{-}th$ ants, correspondingly. As well, $Ant - lion_i^t$ Shows the position of $j\text{-}th$ chosen ant-lion in the $t\text{-}th$ iterations. The equation shows that the ant moves near the preferred ant-lion in a circular cloud described by c and d vectors. For mathematical modeling, the circular cloud radius of arbitrary ant movement is adaptively decreased. The subsequent equation is recommended:

$$c^t = \frac{c^t}{I}\ d^t = \frac{d^t}{I} \tag{3.9}$$

In Eq. (3.9), I denotes a proportion, d^t and c^t Shows the vector encompassing the maximal and minimal value of each variable in $t\text{-}th$ iterations, correspondingly. The last hunting phase is once an ant arrives underneath and is grabbed using its jaw. Then, the ant-lion pulls ants into the soil and eats them. The ant-lion needs to upgrade the position according to the final trap to raise the possibility of capturing novel prey:

$$Ant - lion_j^t = Ant_i^t\ If\ f(Ant_i^t) > f(Ant - lion_j^t) \tag{3.10}$$

In Eq. (3.10), t characterizes the existing iteration, $Ant - lion_i^t$ displays the place of $j\text{-}th$ preferred ant-lion in the $t\text{-}th$ iterations and Ant_i^t Refers to the position of $j\text{-}th$ ant in $t\text{-}th$ iterations.

Here, the optimally chosen ant-lion in every iteration was stored and assumed as an elite. The ant-lion with the highest fitness values can impact the activity of each ant in the iteration. The following equation defines the motion:

$$Ant_i^t = \frac{R_A^t + R_E^t}{2} \tag{3.11}$$

In Eq. (3.11), R_A^t illustrates the arbitrary motion near the chosen ant-lion according to the roulette wheel at $t\text{-}th$ iterationS, R_E^t indicates the arbitrary motion near the elite ant-lion in $t\text{-}th$ iterations, and Ant_i^t Shows the position of $i\text{-}th$ ant in the $t\text{-}th$ iterations. The optimizer added the nature of the ant in hunting the target, and it is given as follows: (1) arbitrary walk, (2) constructing of trap, (3) trapping and in the trap, (4) catching the prey, and (5) rebuilding the trap.

Improved classification performance is the basis for the ALO method's fitness function. It describes a positive integer to characterise the solution candidate's strong performance. One definition of a fitness function in the workplace is the rate at which categorization errors are reduced. A minimal error rate indicates the best solution, whereas a high error rate indicates the worst.

$$\begin{aligned} fitness(x_i) &= ClassifierErrorRate(x_i) \\ &= \frac{number\ of\ misclassified\ samples}{Total\ number\ of\ samples} * 100 \end{aligned} \tag{3.12}$$

3.2.5 LSTM-RNN-based verification process

Finally, the LSTM-RNN model is exploited as a classifier for automated ASR. Generally, a feedforward neural network (FFNN) can be defined in the following [158] [159].

$$Y = F(X, \theta) \tag{3.13}$$

where $X = \{x_1, x_2, \cdots, x_n\}$ refers to an input set, $Y = \{y_1, y_2, \cdots, y_m\}$ represents a collection of outputs, F represents an FFNN module, and θ makes reference to a set of parameters that are associated with the module. Within the context of a classification module, Y stands for a collection of classes. As a kind of FFNN, the Convolutional Neural Network (CNN) is an example. Keyword recognition, semantic segmentation, and picture classification are some of the applications that make use of it. Convolution and pooling layers are part of the CNN method, which sets it apart from other NNs. The convolution layer aims to extract the local feature of the input dataset.

$$Y_F = Conv(X, \theta_{CONV}) \tag{3.14}$$

Conv is one convolution layer, Y_F is a feature subset extracted by the convolution layer from X. θ_{CONV} is a parameter set in the convolution layer. The pooling layer aims to compress the local feature, thus highlighting the feature.

$$Y_{CF} = Poo1(Y_{Fr}\theta_{Pool}) \tag{3.15}$$

The pool is one pooling layer. Y_{CF} represent a set of compressed features, and the fusion of CNN and pooling layer from Y_F is presented. θ_{Pool} is a parameter set in the pooling layer. For a classification module, CNN comprises FC, and $Softmax$ layers are incorporated, and the front of RNN forms a CRNN mechanism. $Y = F(X, \theta)$ categorize features. Here, $Y = F(X, \theta)$ of CNN is $F(\mathrm{X}, \theta)$ of CNN denoted by:

$$Y = Softmax\left(FC\left(Pool\left(Conv(X, \theta_{CONV}), \theta_{pool}\right), \theta_{FC}\right)\right) \tag{3.16}$$

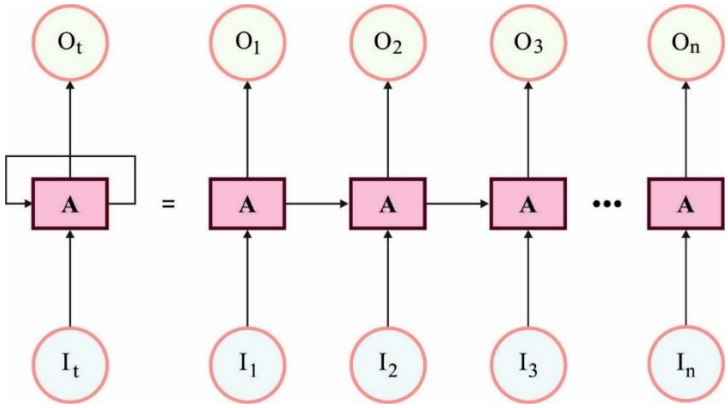

Figure 3.5 Framework of LSTM-RNN

One FC layer that is fully connected is one. One layer of Softmax is referred to as *Softmax*. An alternate form of FFNN is known as RNN. Typically, RNN is used on datasets that have a sequential design, such as those used for voice recognition, machine translation, and other similar tasks. LSTM-RNN is a wise use of RNN that has the potential to overcome the gradient

vanishing issue that occurs with memory cells when it comes to storing long-term data. As a classification method, LSTM-RNN comprises $Softmax$ and FC layers. $y = F(\mathrm{X}, \theta)$ of LSTM-RNN is represented by figure 3.5.

3.3 Results and Discussion

This section examines the speaker verification performance of the TTFEM-AFSV model. The results are inspected under five distinct speakers with 400 samples, as illustrated in Table 1.

Table 3.1 Dataset details

Label	No. of Speakers	No. of Samples
C-1	Speaker-1	80
C-2	Speaker-2	80
C-3	Speaker-3	80
C-4	Speaker-4	80
C-5	Speaker-5	80
Total Number of Samples		**400**

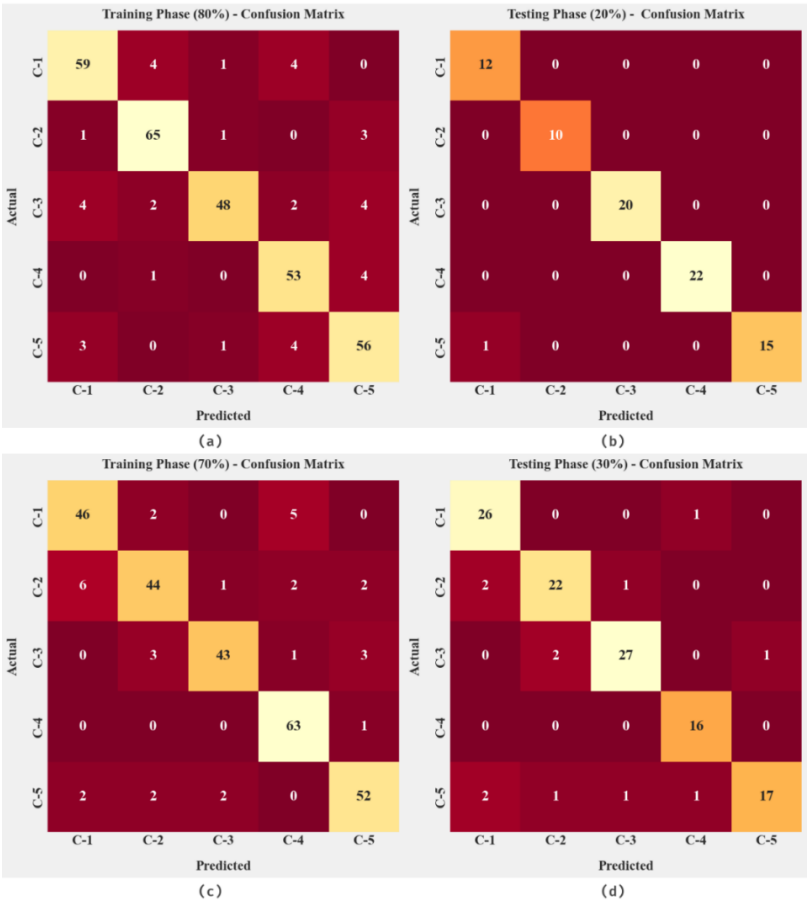

Figure 3.6 Confusion matrices of TTFEM-AFSV approach (a) 80% of TR data, (b) 20% of TS data, (c) 70% of TR data, and (d) 30% of TS data

Figure 3.6 represents the confusion matrices formed by the TTFEM-AFSV model with distinct training (TR) and testing (TS) data sizes. The figure reported that the TTFEM-AFSV model had demonstrated maximum speaker recognition performance.

Table 3.2 provides an overall speaker recognition performance of the TTFEM-AFSV model on 80% of TR data and 20% of TS data. Figure 3.7 illustrates a brief recognition result of the TTFEM-AFSV model on 80% of TR data. The outcomes inferred that the TTFEM-AFSV model recognized all the classes proficiently. For instance, the TTFEM-AFSV model has recognized C-1 samples with $accu_y$ of 94.69%, ER of 5.31%, $prec_n$ of 88.06%, $reca_l$ of 86.76%, F_{score} of 87.41%, and $G_{measure}$ of 87.41%. Meanwhile, the TTFEM-AFSV method has identified C-2 samples with $accu_y$ of 96.25%, ER of 3.75%, $prec_n$ of 90.28%, $reca_l$ of 92.86%, F_{score} of 91.55%, and $G_{measure}$ of 91.56%. Furthermore, the TTFEM-AFSV model has identified C-3 samples with $accu_y$ of 95.31%, ER of 4.69%, $prec_n$ of 94.12%, $reca_l$ of 80.00%, F_{score} of 86.49%, and $G_{measure}$ of 86.77%.

Table 3.2 Result analysis of the TTFEM-AFSV approach with different measures under 80:20 of the TR/TS dataset

Labels	Accuracy	Error Rate	Precision	Recall	F-Score	G-Measure
Training Phase (80%)						
C-1	94.69	05.31	88.06	86.76	87.41	87.41
C-2	96.25	03.75	90.28	92.86	91.55	91.56
C-3	95.31	04.69	94.12	80.00	86.49	86.77
C-4	95.31	04.69	84.13	91.38	87.60	87.68
C-5	94.06	05.94	83.58	87.50	85.50	85.52
Average	**95.12**	**04.87**	**88.03**	**87.70**	**87.71**	**87.79**
Testing Phase (20%)						
C-1	98.75	01.25	92.31	100.00	96.00	96.08
C-2	100.00	00.00	100.00	100.00	100.00	100.00
C-3	100.00	00.00	100.00	100.00	100.00	100.00
C-4	100.00	00.00	100.00	100.00	100.00	100.00
C-5	98.75	01.25	100.00	93.75	96.77	96.82
Average	**99.50**	**00.50**	**98.46**	**98.75**	**98.55**	**98.58**

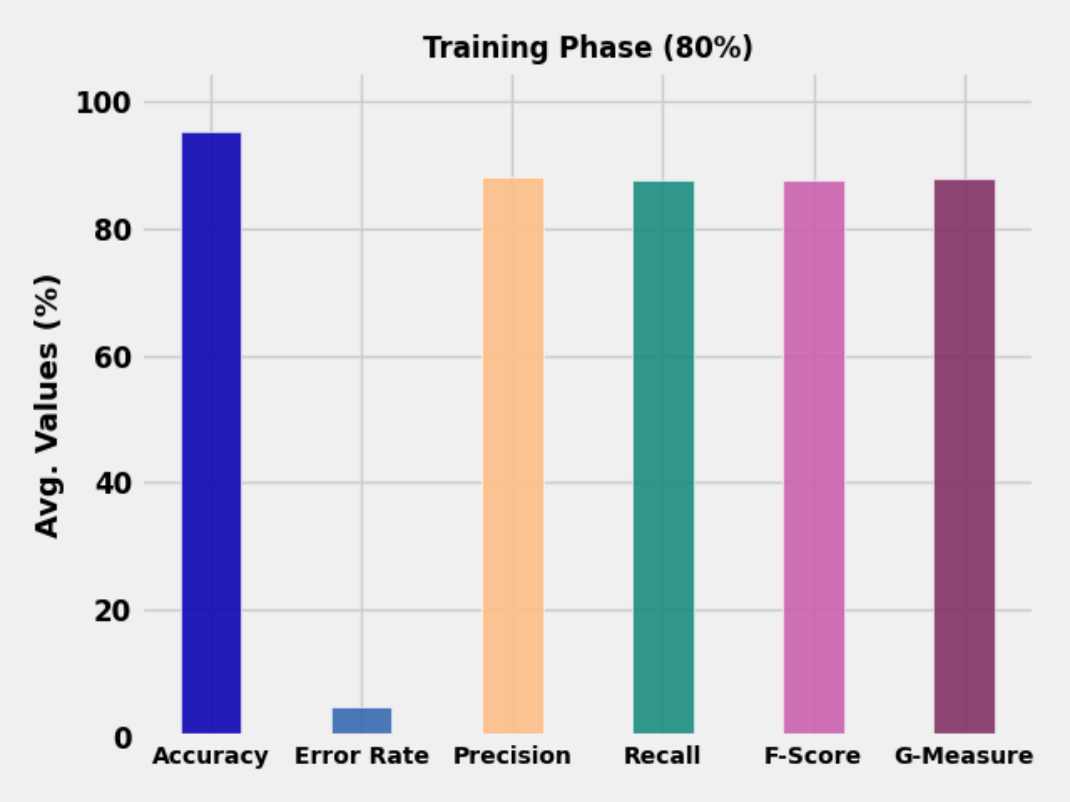

Figure 3.7 Average analysis of TTFEM-AFSV approach under 80% of TR data

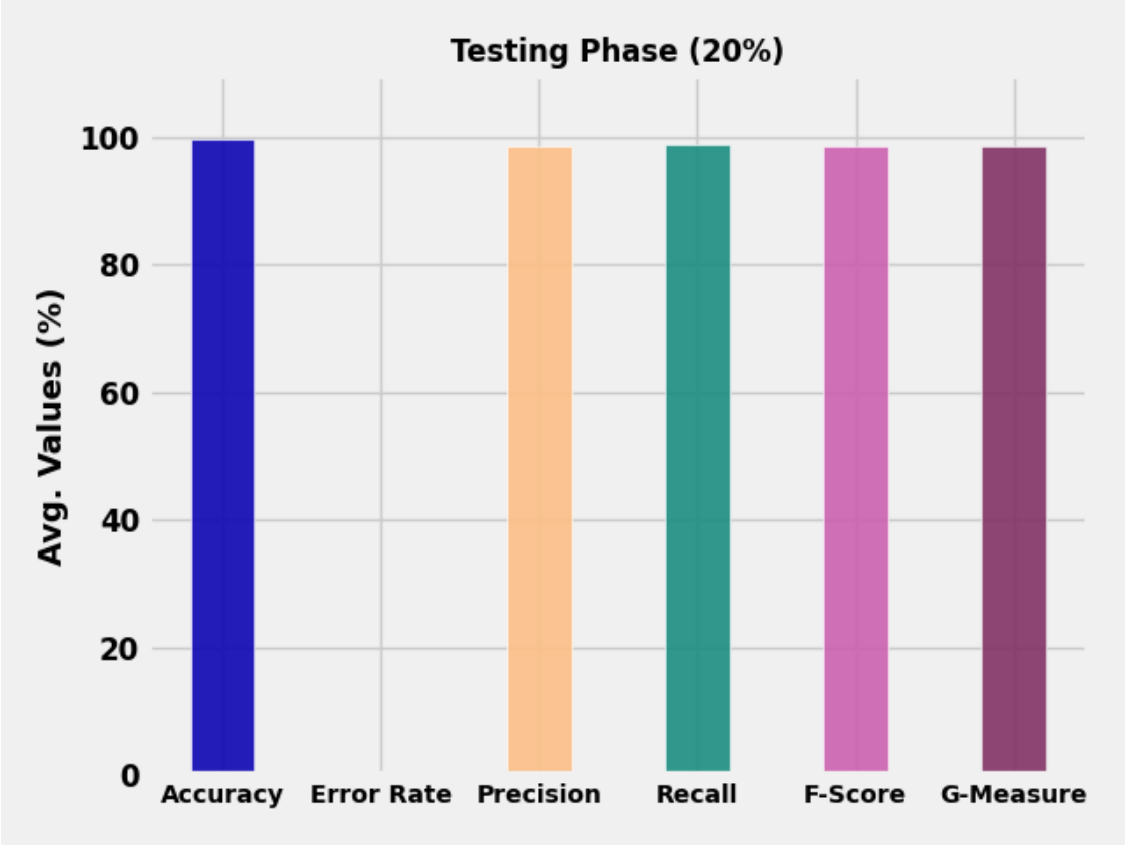

Figure 3.8 Average analysis of TTFEM-AFSV approach under 20% of TS data

Figure 3.8 illustrates a brief recognition outcome of the TTFEM-AFSV model on 20% of the TS dataset. The outcomes inferred that the TTFEM-AFSV approach had identified all the classes proficiently. For instance, the TTFEM-AFSV methodology has identified C-1 samples with $accu_y$ of 98.75%, ER of 1.25%, $prec_n$ of 92.31%, $reca_l$ of 100.00%, F_{score} of 96.00%, and $G_{measure}$ of 96.08%. Meanwhile, the TTFEM-AFSV system has identified C-2 samples with $accu_y$ of 100%, ER of 0%, $prec_n$ of 100%, $reca_l$ of 100%, F_{score} of 100%, and $G_{measure}$ of 100%. Furthermore, the TTFEM-AFSV method has identified C-3 samples with $accu_y$ of 100%, ER of 0%, $prec_n$ of 100%, $reca_l$ of 100%, F_{score} of 100%, and $G_{measure}$ of 100%.

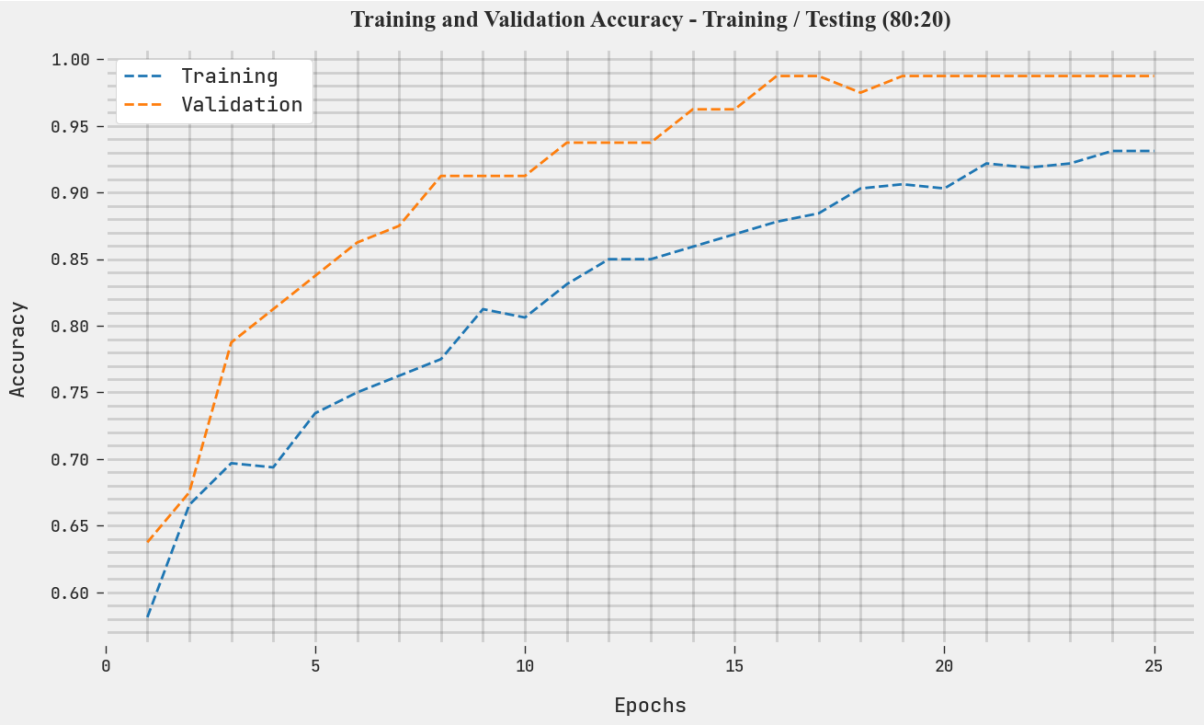

Figure 3.9 TA and VA analysis of TTFEM-AFSV approach under 80:20 of TR/TS data

The training accuracy (TA) and validation accuracy (VA) accomplished by the TTFEM-AFSV approach under 80:20 of TR/TS data is demonstrated in Figure 3.9. The experimental outcomes implied that the TTFEM-AFSV technique has accomplished maximal values of TA and VA. Especially, the VA seemed to be greater than TA.

The training loss (TL) and validation loss (VL) obtained by the TTFEM-AFSV method under 80:20 of TR/TS data are established in Figure 3.10. The experimental outcomes inferred that the TTFEM-AFSV algorithm had achieved minimum values of TL and VL. Particularly, the VL is lesser than TL.

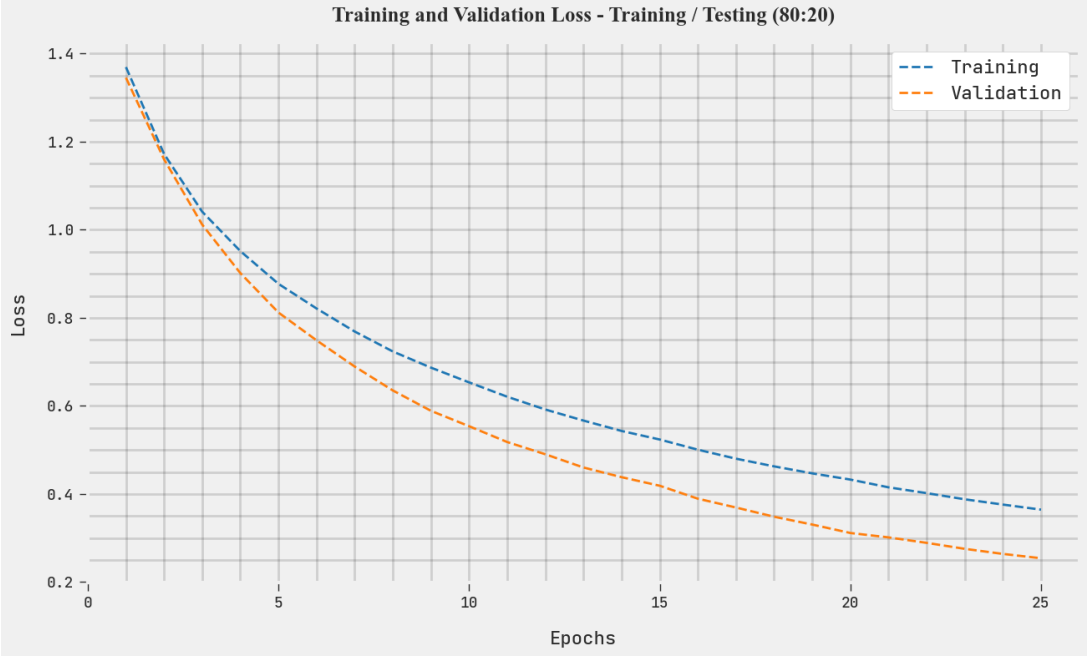

Figure 3.10 TL and VL analysis of TTFEM-AFSV approach under 80:20 of TR/TS data

A clear precision-recall examination of the TTFEM-AFSV technique under 80:20 of TR/TS data is portrayed in Figure 3.11. The figure depicted that the TTFEM-AFSV process has resulted in improved values of precision-recall values under all classes.

A brief ROC analysis of the TTFEM-AFSV approach under 80:20 of TR/TS data is denoted in Figure 3.12. The outcomes indicated the TTFEM-AFSV method had presented its ability to classify distinct classes.

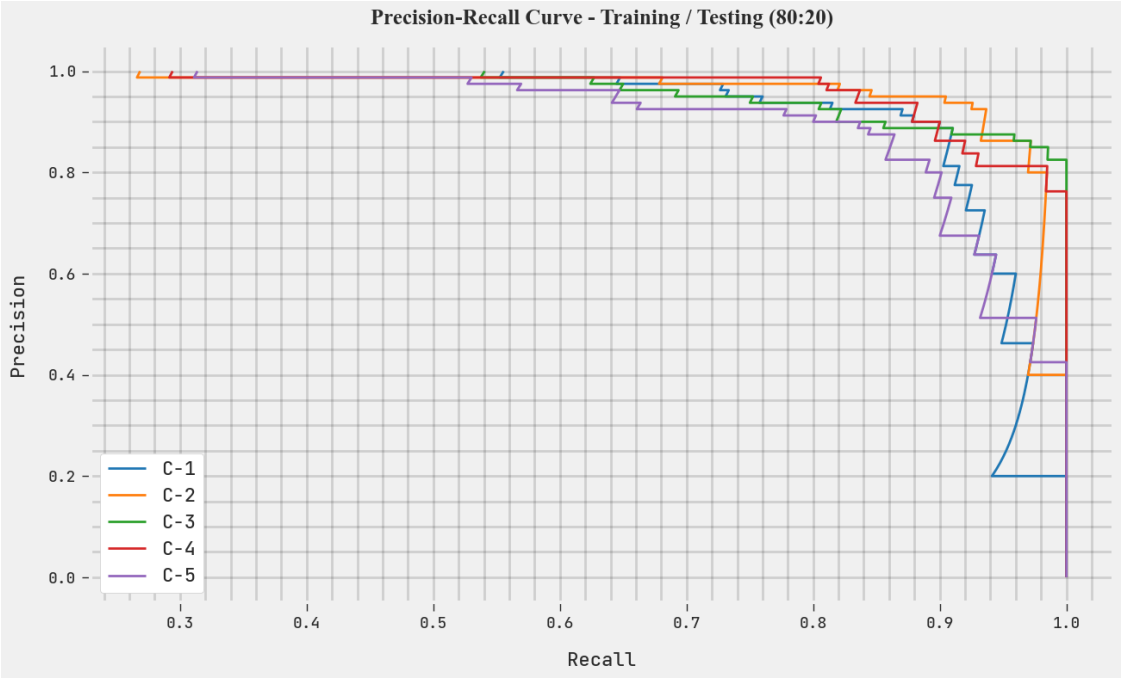

Figure 3.11 Precision-recall analysis of TTFEM-AFSV approach under 80:20 of TR/TS data

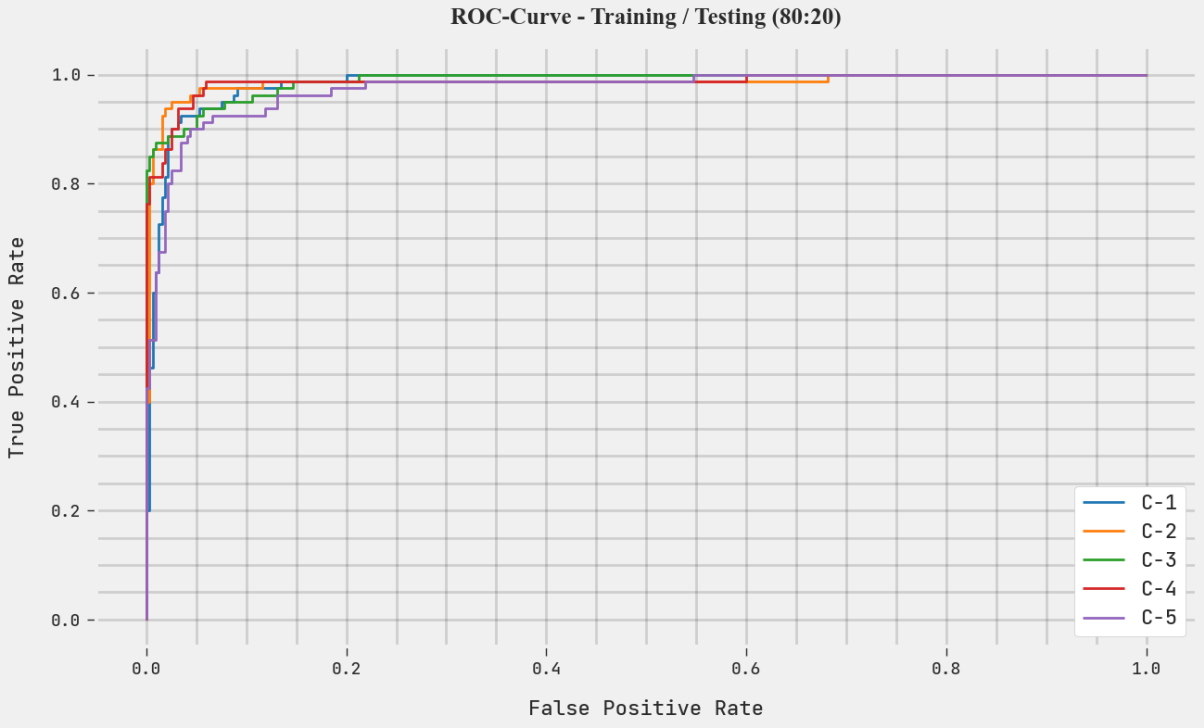

Figure 3.12 ROC analysis of TTFEM-AFSV approach under 80:20 of TR/TS data

Table 3.3 illustrates the overall speaker recognition performance of the TTFEM-AFSV method on 70% of TR data and 30% of TS datasets. Figure. 3.13 inferred a brief recognition outcome of the TTFEM-AFSV approach on 70% of the TR dataset. The result illustrates that the TTFEM-AFSV method has identified all the classes proficiently. For example, the TTFEM-AFSV technique has identified C-1 samples with $accu_y$ of 94.64%, ER of 5.36%, $prec_n$ of 85.19%, $reca_l$ of 86.79%, F_{score} of 85.98%, and $G_{measure}$ of 85.99%. Meanwhile, the TTFEM-AFSV methodology has identified C-2 samples with $accu_y$ of 93.57%, ER of 6.43%, $prec_n$ of 86.27%, $reca_l$ of 80.00%, F_{score} of 83.02%, and $G_{measure}$ of 83.08%. Furthermore, the TTFEM-AFSV system has identified C-3 samples with $accu_y$ of 96.43%, ER of 3.57%, $prec_n$ of 93.48%, $reca_l$ of 86.00%, F_{score} of 89.58%, and $G_{measure}$ of 89.66%.

Table 3.3 Result analysis of the TTFEM-AFSV approach with different measures under 70:30 of the TR/TS dataset

Labels	Accuracy	Error Rate	Precision	Recall	F-Score	G-Measure
Training Phase (70%)						
C-1	94.64	05.36	85.19	86.79	85.98	85.99
C-2	93.57	06.43	86.27	80.00	83.02	83.08
C-3	96.43	03.57	93.48	86.00	89.58	89.66
C-4	96.79	03.21	88.73	98.44	93.33	93.46

C-5	95.71	04.29	89.66	89.66	89.66	89.66
Average	**95.43**	**04.57**	**88.67**	**88.18**	**88.31**	**88.37**
Training Phase (30%)						
C-1	95.83	04.17	86.67	96.30	91.23	91.35
C-2	95.00	05.00	88.00	88.00	88.00	88.00
C-3	95.83	04.17	93.10	90.00	91.53	91.54
C-4	98.33	01.67	88.89	100.00	94.12	94.28
C-5	95.00	05.00	94.44	77.27	85.00	85.43
Average	**96.00**	**04.00**	**90.22**	**90.31**	**89.97**	**90.12**

Figure 3.14 shows brief recognition results of the TTFEM-AFSV system on 30% of the TS dataset. The results inferred that the TTFEM-AFSV technique had identified all the classes proficiently. For example, the TTFEM-AFSV methodology has identified C-1 samples with $accu_y$ of 95.83%, ER of 4.17%, $prec_n$ of 86.67%, $reca_l$ of 96.30%, F_{score} of 91.23%, and $G_{measure}$ of 91.35%. Meanwhile, the TTFEM-AFSV model has identified C-2 samples with $accu_y$ of 100%, ER of 5.00%, $prec_n$ of 95.00%, $reca_l$ of 88.00%, F_{score} of 88.00%, and $G_{measure}$ of 88.00%. Furthermore, the TTFEM-AFSV technique has identified C-3 samples with $accu_y$ of 95.83%, ER of 4.17%, $prec_n$ of 93.10%, $reca_l$ of 90.00%, F_{score} of 91.53%, and $G_{measure}$ of 91.54%.

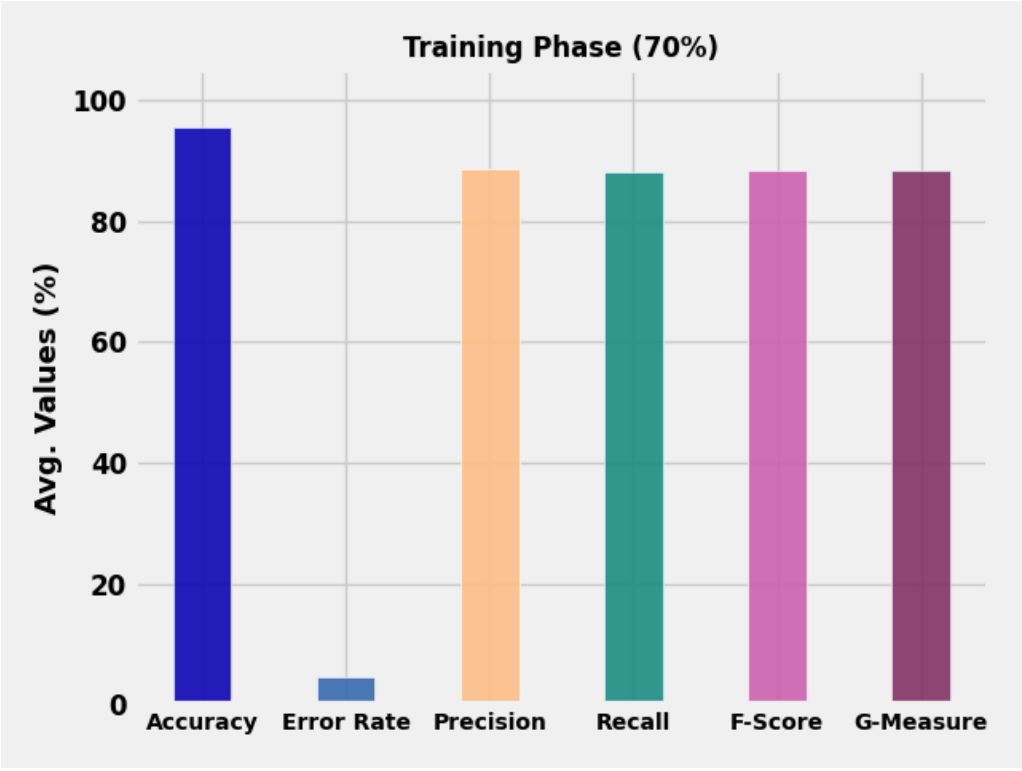

Figure 3.13 Average analysis of TTFEM-AFSV approach under 70% of TR data

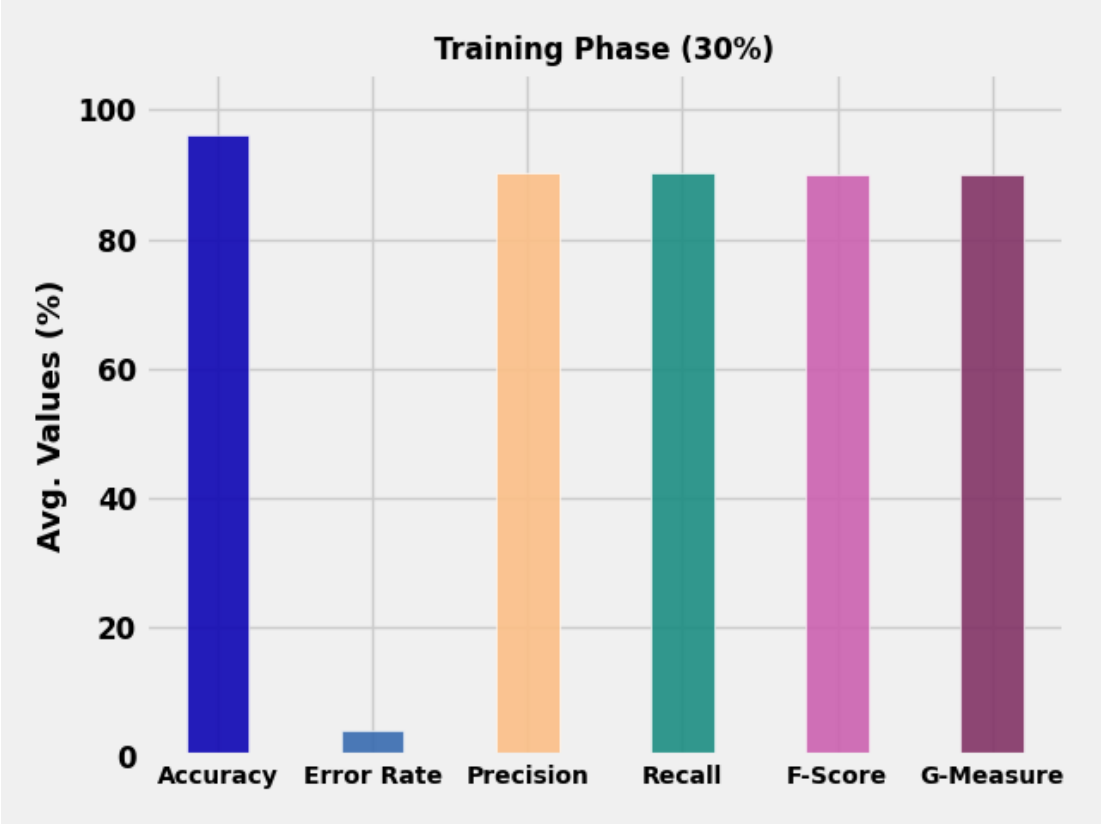

Figure 3.14 Average analysis of TTFEM-AFSV approach under 30% of TS data

Figure 3.15 TA and VA analysis of TTFEM-AFSV approach under 70:30 of TR/TS data

Figure 3.15 shows the TA and VA that were obtained using the TTFEM-AFSV method using 70:30 TR/TS data. Findings from the experiments suggested that the TTFEM-AFSV method had achieved the highest possible TA and VA values. The VA seemed to be higher than the TA, in particular.

Figure 3.16 shows the training loss (TL) and validation loss (VL) that were produced using the TTFEM-AFSV approach with a TR/TS data ratio of 70:30. According to the results of the experiments, the TTFEM-AFSV algorithm got the lowest TL and VL values. The VL is lower than the TL in particular.

Figure 3.16 TL and VL analysis of TTFEM-AFSV approach under 70:30 of TR/TS data

Using 70:30 of TR/TS data, Figure 3.17 shows a clear precision-recall analysis of the TTFEM-AFSV method. Figure 1 shows that across all classes, the TTFEM-AFSV system improved the precision-recall ratio.

A brief ROC analysis of the TTFEM-AFSV technique under 70:30 of TR/TS data is presented in Figure 3.18. According to the findings, the TTFEM-AFSV technique demonstrated its capacity to categorise distinct groups.

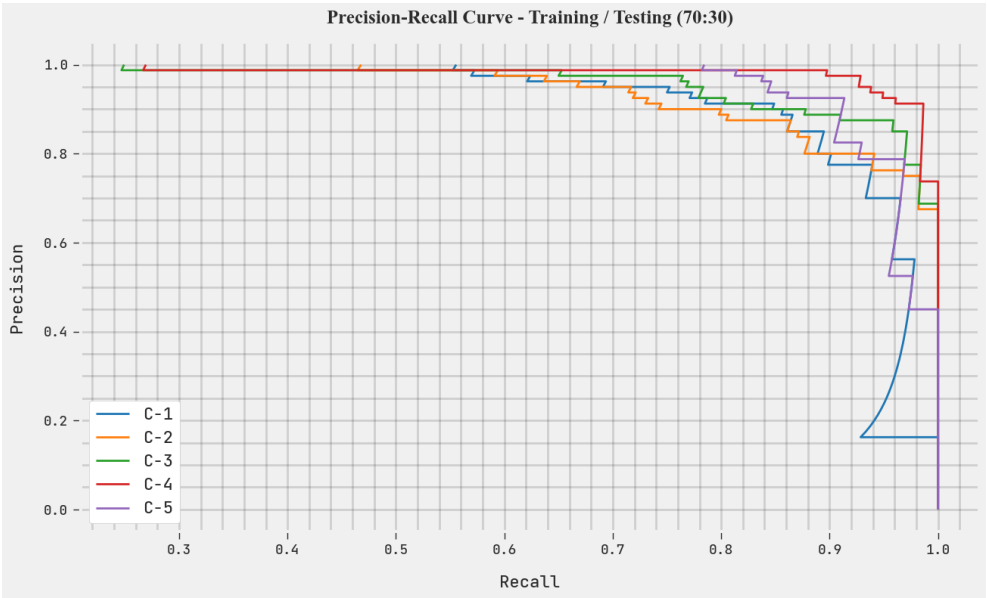

Figure 3.17 Precision-recall analysis of TTFEM-AFSV approach under 70:30 of TR/TS data

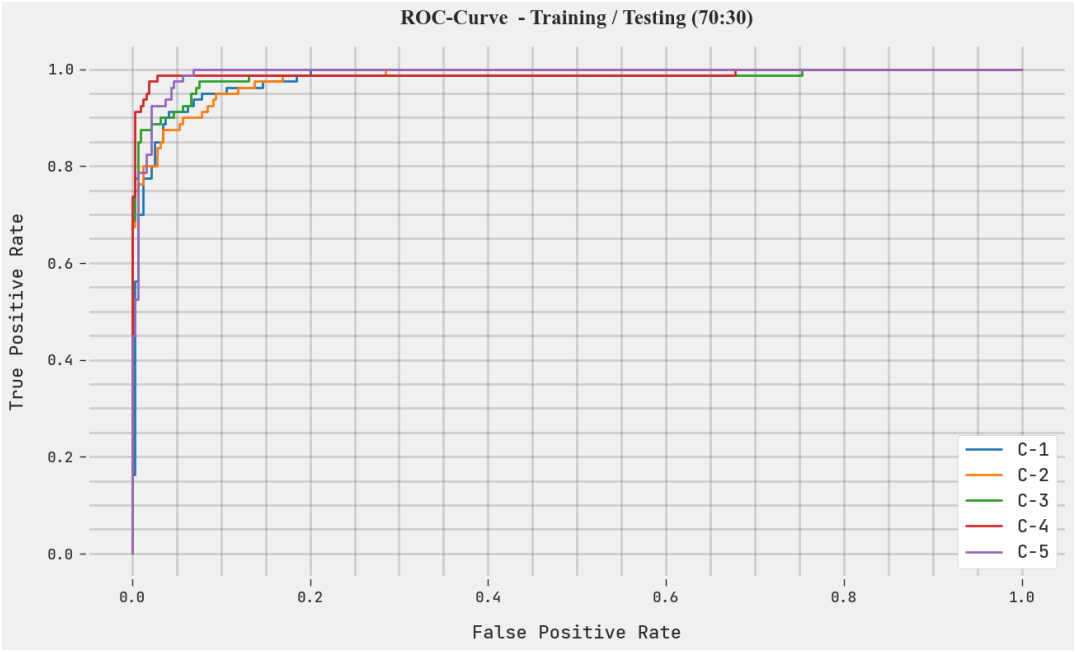

Figure 3.18 ROC analysis of TTFEM-AFSV approach under 70:30 of TR/TS data

To illustrate the TTFEM-AFSV model's improved performance, a comparison study is made in Table 3.4 [152]. Figure 3.19 portrays a comparative $accu_y$ examination of the TTFEM-AFSV model with recent models. The figure implied that the MFCC-SOFM-MLP-GD model had shown lower $accu_y$ of 96.92%. At the same time, the MFCC-SOFM-MLP-GDM, MFCC-SOFM-MLP-BR, MFCC-FW, and fusion models have obtained slightly improved $accu_y$ of 97.05%, 97.62%, 97.32%, and 97.81% respectively. Though the DWT-MFCC technique has resulted in reasonable $accu_y$ of 98.87%, the TTFEM-AFSV model has shown superior results with higher $accu_y$ of 99.50%.

Table 3.4 Comparative analysis of the TTFEM-AFSV approach with recent methodologies

Methods	**Accuracy**	**Error Rate**
TTFEM-AFSV	99.50	0.50
MFCC-SOFM-MLP-GD	96.92	3.08
MFCC-SOFM-MLP-GDM	97.05	2.95
MFCC-SOFM-MLP-BR	97.62	2.38
MFCC-FW	97.32	2.68
DWT-MFCC	98.87	1.13
FUSION	97.81	2.19

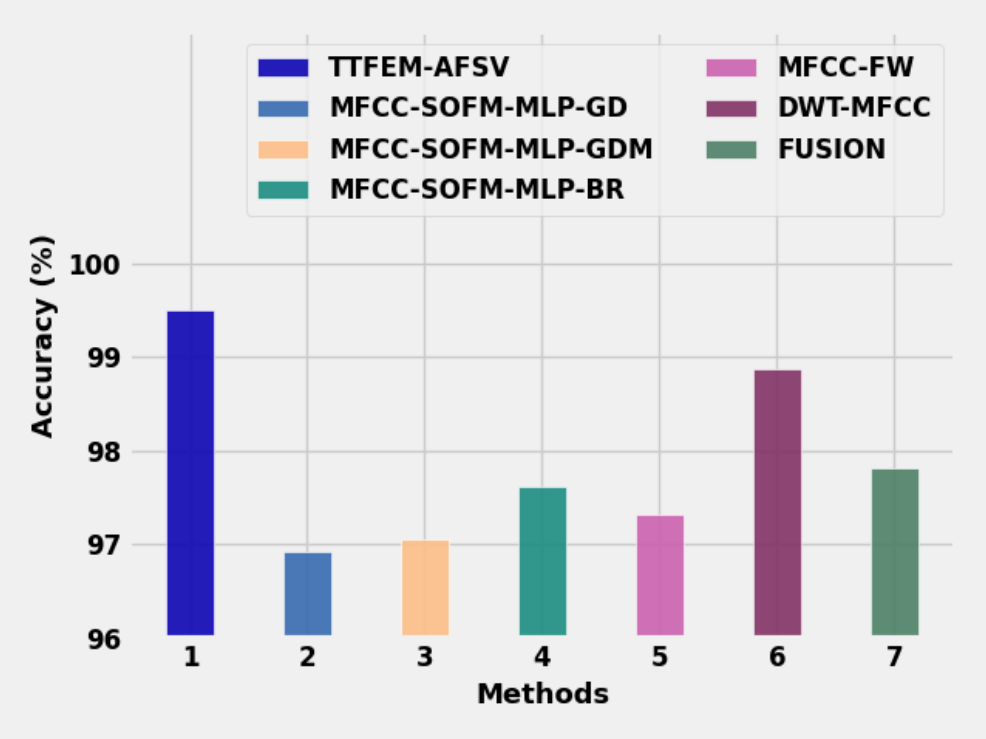

Figure 3.19 $Accu_y$ analysis of the TTFEM-AFSV approach with recent methodologies

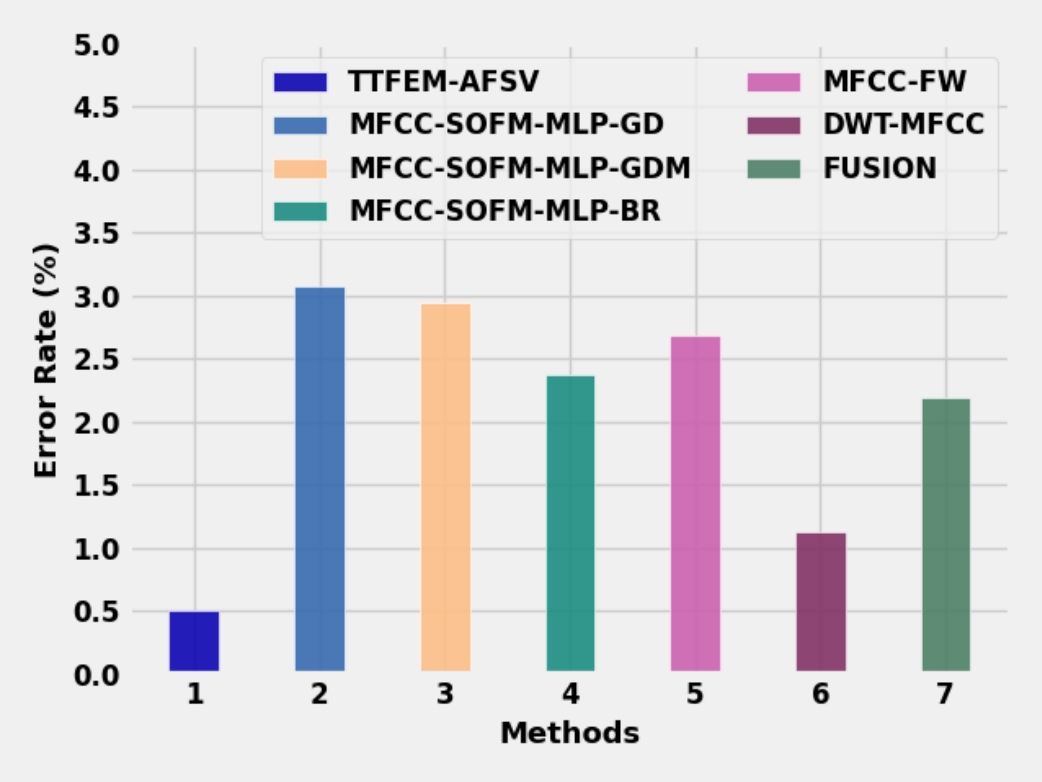

Figure 3.20 Error rate analysis of TTFEM-AFSV approach with recent methodologies

Figure 3.20 depicts a comparative error rate (ER) examination of the TTFEM-AFSV method with current models. The figure shows that the MFCC-SOFM-MLP-GD method has a greater

ER of 3.08%. At the same time, the MFCC-SOFM-MLP-GDM, MFCC-SOFM-MLP-BR, MFCC-FW, and fusion approaches have attained slightly reduced ER of 2.95%, 2.38%, 2.68%, and 2.19% correspondingly. Though the DWT-MFCC method has resulted in a reasonable ER of 1.13%, the TTFEM-AFSV system has illustrated greater outcomes with the least ER of 0.50%. Therefore, the TTFEM-AFSV model has shown effective speaker verification performance over other models.

3.4 Applications

Speaker Authentication: The two-tier feature extraction approach is very appropriate for speaker authentication in forensic investigations. The system may authenticate the asserted identity of a speaker by comparing the retrieved characteristics from a voice sample with pre-registered speaker models.

Voice Evidence Analysis: This approach aids in the examination and verification of speaker identities in audio recordings, particularly in forensic contexts where such recordings may serve as pivotal evidence. It assists in ascertaining if a certain voice corresponds to the stated identity or reveals any discrepancies.

Forensic Speaker Profiling: The retrieved attributes may aid in the creation of comprehensive profiles of speakers, including distinct speech characteristics. Profiling may be advantageous in criminal investigations, offering further insights on suspects or individuals of interest.

3.5 Challenges and Considerations

Data variability: Forensic applications often require handling heterogeneous and turbulent information. The model must possess sufficient robustness to effectively manage fluctuations in recording settings, background noise, and emotional states.

Ethical and Legal Implications: Forensic speaker verification raises ethical and legal considerations, especially regarding the use of voice evidence in legal proceedings. Adherence to privacy and legal standards is crucial.

3.6 Summary

Forensic speaker verification using a newly-developed TTFEM-AFSV model is detailed in this chapter. To begin with, the TTFEM-AFSV model removed the background noise from speech signals using the AMF method. A further step involves feeding the MFCC and spectrograms into the Inception v3 model. Afterwards, the Inception v3 model's hyperparameters are optimised using the ALO approach. The last step in automated ASR is the use of the LSTM-RNN model as a classifier. A battery of experiments is used to validate the TTFEM-AFSV model's performance. The comparison shows that the TTFEM-AFSV model outperforms the newer approaches. As a result, forensic applications that need efficient real-time ASR may find the TTFEM-AFSV model useful. To further improve the TTFEM-AFSV model's overall performance, future work may include developing a mix of fusion-based deep learning models.

4. Chapter - Speaker Diarization

Automated speech processing techniques like speaker diarization use segmentation and tagging of audio recordings to separate voices and attribute specific parts of speech to them. Identifying the speakers and their respective times in an audio recording from a lack of background information is the main objective.

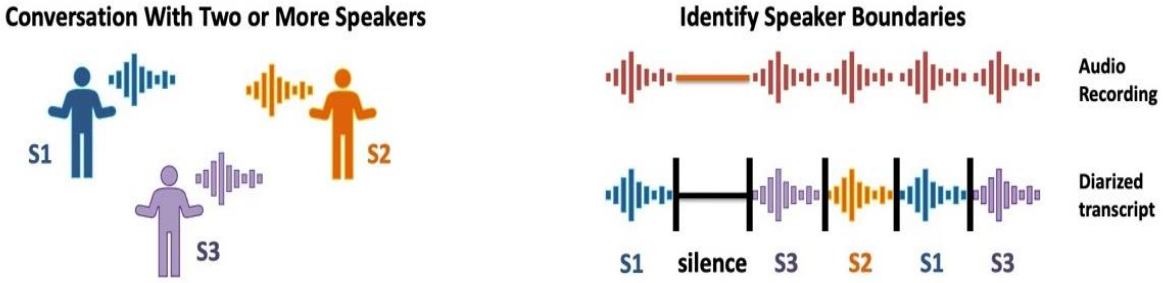

Figure 4.1 Speaker Diarization

Speaker diarization involves annotating an input audio stream by assigning distinct sources to overlapping temporal areas of the sound. Speakers, songs, ambient noise, and other aspects of the signal's channel or origin could be considered sources [160]. The speaker diarization mechanism is shown in Figure 4.1. Speaker diarization is the term used when the stated classes match the identities of the various speakers in the recording. One often seen definition in the literature is that of providing a response to the query "who spoke when". When an audio stream is divided into segments that include a single speaker and then organised according to the speakers' identities, this process is called speaker diarization. A speaker diarization system's output is often a set of time-based segments, whereby a single speaker's voice is featured in each segment [161] [162]. Each segment is assigned an abstract ID that identifies the speaker who is speaking in that particular segment. Many crucial applications rely on the technique, including forensic analysis, voice-controlled devices, and transcribing services.

Speaker diarization may be seen as the merging of speaker segmentation and speaker clustering tasks. Speaker segmentation is the process of identifying possible locations in an audio stream where there may be a shift in speakers [163]. Speaker clustering, on the other hand, involves grouping these identified segments into clusters based on the speakers. Speaker diarization is a job that operates on two assumptions:

a) It is unknown in advance how many speakers will be participating
b) The names of the speakers in the recording are also unknown

It is an unsupervised job in which the system must predict the precise number of speakers and appropriate speaker models.

4.1 Speaker diarization process

Figure 4.2 represents a typical speaker diarization pipeline. There are five primary stages, outlined as follows:

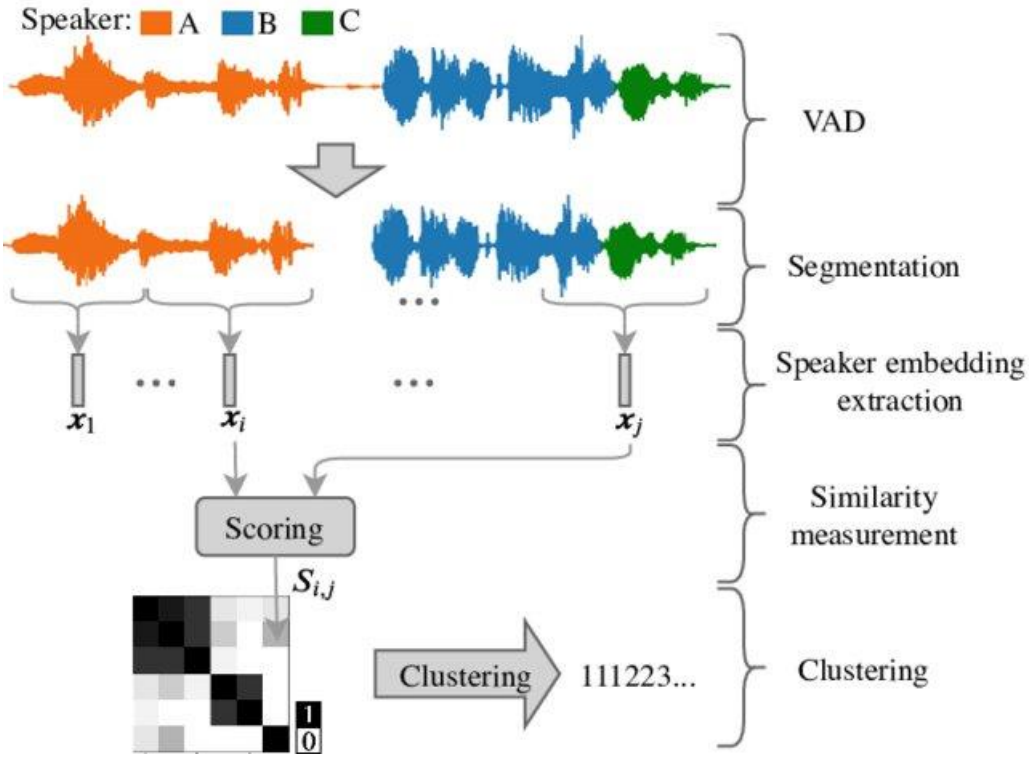

Figure 4.2 A typical speaker diarization pipeline [164]

- Voice Activity Detection (VAD): Distinguishing speech from other noises, such as stillness, is the main goal of voice activity detection. For the final VAD segmentation, the speech segments' borders should be exact and accurate.
- Speech Segmentation: Subsequently, the audio file is sometimes segmented according to pauses or some other acoustic characteristic.
- Feature Extraction: From every segment, acoustic characteristics including pitch and Mel-Frequency Cepstral Coefficients (MFCCs) are retrieved.
- Clustering: The extracted features are grouped into clusters using clustering algorithms like k-means or Gaussian Mixture Models (GMMs). Each cluster represents the speech of a particular speaker.
- Labelling: The identified clusters are labelled with speaker IDs, creating a speaker timeline or diarization output.

4.1.1 Challenges in Speaker Diarization

a. Overlapping Speech: When multiple speakers talk simultaneously, distinguishing their speech becomes challenging. Overlapping speech can lead to errors in cluster assignments
b. Speaker Variability: Variability in a speaker's voice due to factors like emotion, health conditions, or different recording environments can affect diarization accuracy
c. Lack of Training Data: Limited training data for specific speakers or diverse acoustic conditions can hinder the performance of diarization systems, especially in real-world scenarios
d. Non-speech Sounds: Background noises, music, or non-speech sounds may be present in the audio, complicating the segmentation and clustering process
e. Speaker Turn Changes: Quick changes in speaker turns without significant pauses may result in errors in segmentation and misattribution of speech segments
f. Data Length: Short audio recordings might not provide sufficient context for accurate speaker diarization, especially when dealing with multiple speakers
g. Scalability: The diarization job becomes more hard as the audio recording time or number of speakers grows
h. Cross-Language Variability: Variations in pronunciation, accent, and language can pose challenges, particularly in multilingual settings
i. Real-time Processing: Achieving real-time speaker diarization is a demanding task, requiring efficient algorithms to process audio streams promptly
j. Evaluation Metrics: The absence of standardized evaluation metrics can make it challenging to objectively compare the performance of different diarization systems

Research continues to address these challenges, making speaker diarization a dynamic and evolving field in automatic speech processing. To design a speaker diarization system that is both quick and accurate over the course of several sessions is the fundamental purpose of this chapter. While mass meetings and communication proliferate, speakers' diarization can be burdened to enhance the readability of speech transcripts. To address these issues, this study develops an automated speaker diarization system using an arithmetic optimization algorithm with a deep belief network (ASDS-AOADBN) technique.

4.2 Speaker diarization model framework

An ASDS-AOADBN system for the automatic speaker diarizaion process, which accurately identifies and classifies the speaker signal from the input audio signals is developed. The

ASDS-AOADBN technique comprises SGF-based preprocessing, MFCC feature extraction, DBN-based classification, and AOA-based hyperparameter tuning to accomplish this. The ASDS-AOADBN system's major purpose lies in identifying and classifying speaker signals from input audio signals. To accomplish this, the proposed ASDS-AOADBN method first improves the quality of the audio input signals by removing noise using the Savitzky-Golay filter (SGF) approach. To top it all off, the preprocessed audio signals include Mel Frequency Cepstral Coefficient (MFCC) characteristics extracted from them. In order to identify speakers, the ASDS-AOADBN method employs the DBN model, which reliably ascertains the speaker's class label. To improve the ASDS-AOADBN algorithm's detection performance, the AOA-based hyperparameter tuning procedure is carried out.

4.2.1 State of art

The speaker diarization system finds homogeneous speaker areas using unsupervised or supervised techniques. The classical speaker diarization system is unsupervised, in which limited data is generally available, and the number of speakers is unknown [165]. Lately, many supervised methods have been advanced where the whole diarization pipeline has been trained for clustering and segmentation. The data for diarization involves meeting recordings, TV/ talk shows, and telephone [166]. Implementing diarization on audio-only recordings was complicated because of numerous factors, including reverberations, environmental noise, short utterances, and overlapping speech [167]. Audio-visual techniques have been devised to assist the diarization task for robust clustering and segmentation of homogeneous speakers. Such methods generally find the active speaker in the video domain by implementing a motion detection approach on the face [168]. Multimodal diarization methods are devised based on available datasets with dissimilar audio recording configurations: headset or lapel, single omnidirectional microphone, and microphone array [169]. Likewise, for recording video, individual camera streams or single wide-angle cameras for all speakers may be accessible.

Further, speakers may be moving or static in multiparty interaction [170]. Classical methods for speaker diarization utilize the Gaussian Mixture model (GMM) or Hidden Markov model (HMM) with an agglomerative hierarchical clustering (AHC) system for clustering and unsupervised segmentation [171]. This method initialized many clusters and hierarchically integrated them as per Bayesian data conditions.

While the diarization system does not necessitate speaker enrolment, they can just make abstract labels of active speakers in an audio segment. Then, the speaker identification system

offers non-abstract labels through the enrolment of the participating speakers [172]. It is important to develop diarization techniques that are not required to pre-enroll the speakers [173]. Numerous speech-based applications, such as authentication, identification, and recognition, rely on the processing and analysis of human voice signals to identify and quantify certain physical properties. Artificial Neural Network (ANN) is a useful computational tool [174]. Utilizing deep learning (DL) techniques, ANNs try to mimic the human brain's performance to execute the functionalities indulged in speech processing and enhance the outcomes [175].

4.2.2 Chapter contributions

- An automated speaker diarization system using an arithmetic optimization algorithm with a deep belief network (ASDS-AOADBN) technique
- The presented ASDS-AOADBN technique for preprocessing uses Savitzky–Golay filter (SGF) technique
- The preprocessed audio signals undergo feature extraction to get Mel Frequency Cepstral Coefficient (MFCC) characteristics
- The ASDS-AOADBN technique identifies speaker signals using a DBN model that correctly determines the speaker's class label
- To enhance the DBN model's performance, the AOA-based hyperparameter tweaking procedure is carried out. On the DIHARD and CALLHOME datasets, the ASDS-AOADBN algorithm produces a thorough experimental result

4.3 ASDS-AOADBN model

A novel ASDS-AOADBN system is created for the automated speaker detection procedure. It successfully detects and categorises the speaker signal from the input audio signals. This is achieved using the ASDS-AOADBN technique's SGF-based preprocessing, MFCC feature extraction, DBN-based classification, and AOA-based hyperparameter tuning. Figure 4.3 represents the overall flow of the ASDS-AOADBN algorithm.

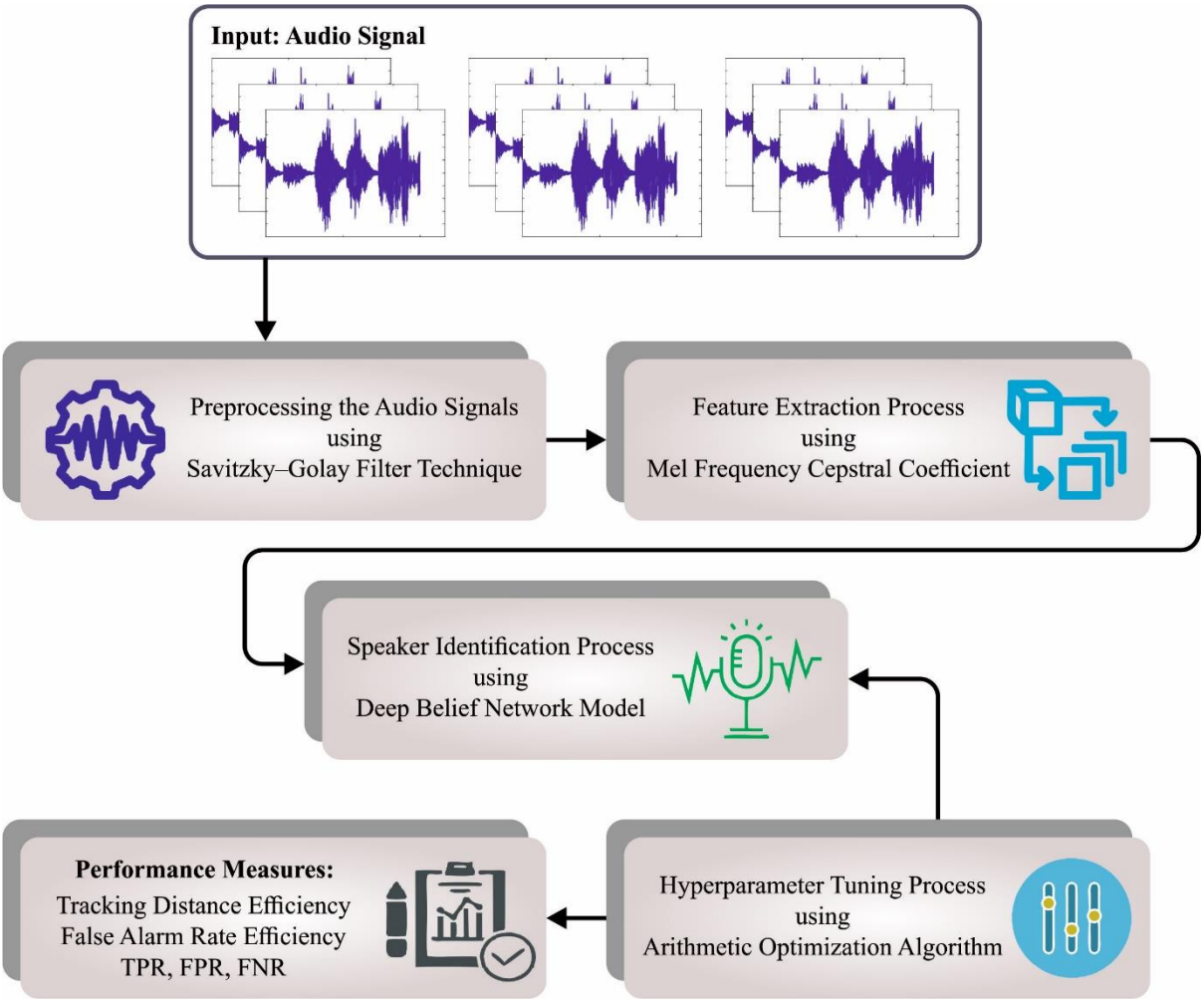

Figure 4.3 Overall flow of the ASDS-AOADBN approach

4.3.1. SGF-based Preprocessing

The proposed ASDS-AOADBN technique primarily pre-processed the audio input signals using the SGF technique. Preprocessing of the speech signal plays a vital role in the design of the ASR model [176]. The SGF method was used to remove the white noise from the input speech signal to improve the accuracy and outcome of the ASR model. It chooses the accurate frame size and signal relativity and demands repeated observation. In general, the SGF method exploits linear least squares to smoothen the data for generating a maximum PSNR and setting the initial state of the signal. The SGF method recovers the real structure of the signal through noise removal. The Savitzky–Golay channel is interconnected with the series of current details focused on raising the SNR without signal distortion [109]. The subset of consecutive data points is achieved by using a low-request polynomial with a linear least square model and convolutional of the substantial amount of polynomial. The details hold a collection of n$\{x_i, y_j\}$ points, where $j = 1,2 \ldots n$, and x shows the independent parameter, whereas y is detected

esteem, vocalized with the set of m convolution coefficients, C_i, and can be represented as follows:

$$Y_j = \sum_{i=-(m-1)/2}^{i=(m-1)/2} C_i y_{j+i} \frac{m+1}{2} \leq j \leq n - \frac{m-1}{2} \quad (4.1)$$

The SGF method requires three diverse resources of data: the size of the frame (f), noisy signal (x), and order of the polynomial (k). The better-fit estimation of k and f for the signal can be determined mainly through using a trial and error algorithm.

Pre-processing with the Savitzky-Golay filter (SGF) is necessary to improve the signal-to-noise ratio in audio data. SGF is perfect for noise reduction since it smoothly enhances the input signal while maintaining its essential characteristics. SGF preprocessing provides the Deep Belief Network Model and the Arithmetic Optimization Algorithm with cleaner and more informative audio input. It improves the diarization system's overall performance by increasing its accuracy and lowering false positives. It was selected in order to maximize the model's performance by tackling background noise, which is one of the main obstacles to speaker diarization.

4.3.2. MFCC Features

The MFCC features are derived from the preprocessed audio input signals at this stage. The feature extraction with MFCC is highly known in speaker recognition, which gives the purpose of recognizing speakers dependent upon low-level properties [177]. Initially, the extraction makes adequate data for speaker discernment and captures these data in form and size, allowing effective modelling. Thus, feature extraction is a data reduction technique that extracts speakers' essential characteristics at a reduced data rate. The extraction feature takes the digital voice signals and turns them into feature vectors, which are a collection of mathematical descriptors. Some methods for extracting features are used by speaker detection systems. A popular way to obtain the MFCC from speech signals is by using cepstral analysis. Following the input signal's framing and windowing, the amplitude of the output spectrum may be distorted by mel-scale after the Fourier transform has been considered. N subsamples make up each subsample of the 1D signal. This section is named Frames, and the inspiration behind it is the quasistationary nature of the 1D signal. But if the signal is inspected over the distinct section, which is satisfactorily shorter in duration, it can be regarded as stationary and exhibits stable features. Frame overlap prevents loss of information. All the frames begin at a particular offset of L samples concerning the prior frames where LBN. For every frame, a windowing

function is commonly used for increasing the continuity among neighboring structures. The windowing function includes the Hamming, rectangular, flattop, and Blackman windows. Now, the magnitude spectrum $|X(k)|$ was scaled in magnitude and frequency. Firstly, the frequency is logarithmically scaled employing Mel filter bank (k, m), and later the logarithm was obtained as follows:

$$X'(m) = \ln \left| \sum X\,(k) H(k, m) \right| \tag{4.2}$$

For $m = 1, 2$, M; where m denotes the number of filter banks.

The Mel filter bank is a set of triangular filters determined by center frequency evaluated on the Mel scale. The filter counts are the parameter that affects the detection performance of the model. Lastly, the MFCC is attained by calculating the DCT of $X'(m)$ as follows:

$$C_l = \sum X'(m) \cos |l. - M'' p| m - \wedge 1| \,| \tag{4.3}$$

for$1 = 1, 2, \ldots, M$, whereas C_l. is the l_{th} MFCC.

Delta is the name of the MFCC feature vector's first-order regression coefficient. Delta-Delta refers to the 2nd-order regression coefficient.

The count of the resulting MFCCs is chosen between 12 and 20, after which the first few coefficients of the signal data are identified. The average log energy of frames is indicated by the 0^th co-efficient. The experience with voice identification demonstrates the benefits of using delta and delta-delta coefficients that lower the word rate of errors. The data redundancy of elements in the feature vector rises; however, the original group of characteristics of the MFCC is relatively connected, followed by the addition of delta and delta-delta characteristics.

4.3.3. Speaker Diarization using the DBN model

For automated speaker diarization, the DBN model is utilized. The RBM is an undirected, bipartite graphical model that includes the hidden layer (output) and visible layer (input) [178]. Both layers are composed of m hidden units and n visible units correspondingly, and there is a bias in every unit. Besides, there is no connection within visible or hidden states. Figure 4.4 demonstrates the architecture of classical DBN.

The number of layers/units (RBMs): The visible layer and the hidden layer are the two layers that make up BMs in general. Additional important deliberation is the number of units in each layer.

An overview of how to set the number of units and layers in an RBM is provided below:

- *Visible Layer:* The input data is signified by the visible layer. The number of features or variables in your input dataset should match the number of units in the visible layer.
- *Hidden Layer:* The intricate connections between the features in the data are captured by the hidden layer. A hyper-parameter that must be selected for challenge and the intricacy of the patterns want the RBM to identify is the number of units in the hidden layer. When using RBMs, the model architecture is often symmetric, which means that the quantity of units in the visible and hidden layers is the same.

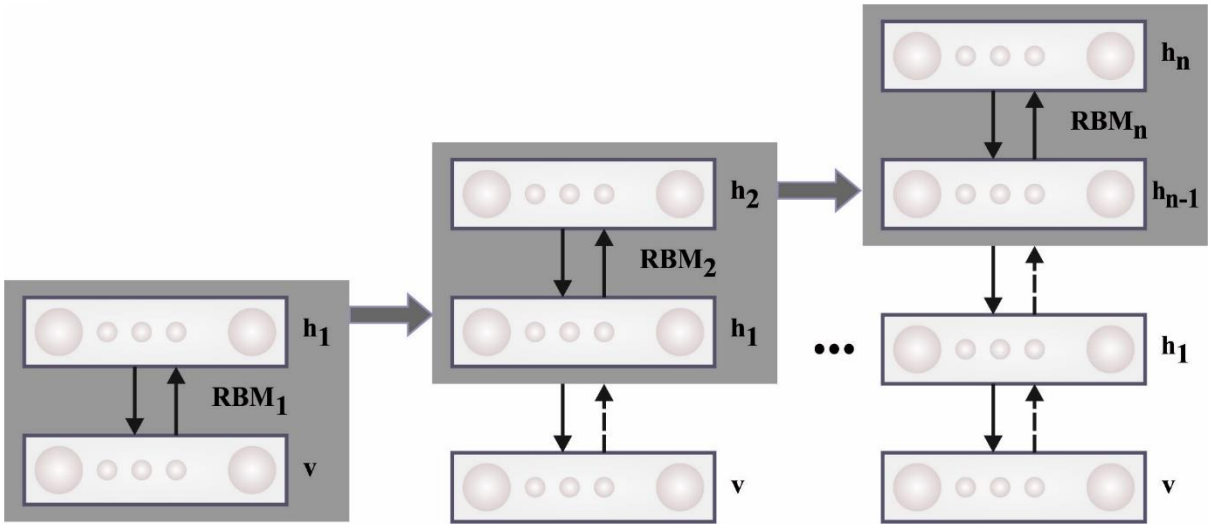

Figure 4.4 Architecture of DBN

The activation probability of j^{th} hidden units are evaluated by Eq. (4.4):

$$p(h_j = 1|v) = \sigma\left(b_j + \sum_{i=1}^{n} v_i\, w_{ij}\right), \tag{4.4}$$

Now, b_j denotes the bias of j^{th} hidden units, $\sigma(\cdot)$ indicates the sigmoid function, and w_{ij} represent the weight connection between i^{th} and j^{th} visible and hidden units.

As well, the activation probability of i^{th} visible units are evaluated using Eq. (4.5) if the hidden vector $h(h_1, \dots, h_j, \dots, h_m)$ is known,

$$p(v_i = 1|h) = \sigma\left(a_j + \sum_{j=1}^{m} h_j\, w_{ij}\right), \tag{4.5}$$

For $i = 1,2, \dots, n$, where a_j denotes the bias of i^{th} visible units.

A contrastive divergence (CD) model is used to improvise the RBM. The model-based iterative learning procedure of RBM for binomial units is shown below.

Step 1: Initialize the parameters: the amount of training data N, the number of visible units n, the number of hidden units m, the visual bias vector a, the weighting matrix W, learning rate

e, and the hidden bias vector b.

Step 2: Allocate the sample x in the trained dataset that is the primary state v_0 of the visible layer.

Step 3: Where $j = 1,2, \dots, m$, compute $p(h_{0j} = 1|v_0)$ based on Eq. (4), and extract $h_{0j} \in \{0,1\}$ from the $p\left(h_{0j} = 1\middle|v_0\right)$ conditional distribution.

Step 4: Where $i = 1,2, \dots, n$, compute $p(v_{1i} = 1|h_0)$ based on Eq. (5), and extract $v_{1i} \in \{0,1\}$ from the $p(v_{1i} = 1|h_0)$ conditional distribution.

Step 5: Compute $p(h_{1j} = 1|v_1)$ based on Eq. (4.4).

Step 6: Upgrade the parameter using the subsequent formula:

$$W = W + \epsilon(p(h_0 = 1|v_0)v_0^{\mathrm{T}} - p(h_1 = 1|v_1)v_1^{\mathrm{T}}),$$

$$a = a + \epsilon(v_0 - v_1),$$

$$b = b + \epsilon(p(h_0 = 1|v_0) - p(h_1 = 1|v_1)).$$

Step 7: Allocate additional samples from the training dataset that the primary state v_0 of visible layer, and repeat Steps 3-7 still the N trained dataset is attained.

The structure of DBN with k^{th} hidden layers along with the layer-by-layer pre-training model. The activation of k^{th} hidden state about x input instance is evaluated as follows:

$$A_k(x) = \sigma\left(b_k + W_k\sigma\left(\cdots + W_2\sigma(b_1 + W_1x)\right)\right), \tag{4.6}$$

In Eq. (4.6), W_u and $b_u(u = 1,2, \dots, k)$ are, correspondingly, the weighted matrix and hidden bias vector of u^{th} RBM. In addition, σ denotes the logistic sigmoid function $(x) = 1/(1 + e^{-x})$. The DBN exploits deep structure to attain the best feature representation and implements the layer-by-layer pre-training to improvise the inter-layer weighted matrices.

The uses of DBN:

- DBN can be used to divide audio recordings into separate speaker segments in an efficient method. Accurate segmentation is enabled by the model's ability to recognize difficult patterns in the auditory characteristics connected to distinct speakers.
- Learning hierarchical feature representations is an area in which DBN excels. It removes the requirement for human feature engineering in speaker diarization by automatically extracting pertinent characteristics from unprocessed audio signals.
- DBN is accomplished by producing strong speaker embedding that captures the distinctive qualities of each speaker. For subsequent tasks like speaker identification or emotion detection, these embeddings may be useful.

- The DBN model may lessen the requirement for laborious physical annotation of training data because it can learn from unlabelled data. In the process of developing a speaker diarization system, this can save time and resources.

4.3.4 Hyperparameter tuning using AOA

Lastly, the AOA-based hyperparameter tuning process is performed to adjust the hyperparameter values of the DBN technique. Abualigah *et al.* [179] presented a new population-based metaheuristic algorithm termed AOA that exploits the distribution features of 4 most critical arithmetic operators in mathematics, such as Subtraction ($S\varepsilon - \varepsilon$), Division ($D\varepsilon \div \varepsilon$), Addition ($A\varepsilon + \varepsilon$), and Multiplication ($M\varepsilon \times \varepsilon$). In AOA, a primary population can be created arbitrarily by utilizing below formula:

$$x = LB + (U - L) \times \partial \tag{4.7}$$

whereas x denotes the population solution. The upper and lower boundaries of searching space are represented as U and L, correspondingly. ∂ signifies the random parameter that takes a value among. Previously, at the beginning of AOA, the choice of exploitation and exploration was executed depending on the Math Optimizer Accelerated (MOA) function that can be evaluated by Eq. (4.8).

$$MOA(C\_Iter) = \text{Min} + C\_Iter \times \left(\frac{\text{Max} - \text{Min}}{M\_Iter}\right) \tag{4.8}$$

whereas $MOA(C\_Iter)$ implies the functional outcome at t^{th} iteration. C_Iter represents the current iteration, which lies within one and maximal iteration count (M_Iter). Min and Max represent the minimum and maximum of MOA.

The exploration of global searching in AOA search can be executed by utilizing Multiplication (M) and Division (D) operators-based searching approaches that are expressed in Eq. (4.9).

$$x_{i,j}(t+1) = \begin{cases} best\left(x_j\right) \div (MoPr + \epsilon) \times \left(\left(U_j - L_j\right) \times \mu + L_j\right), rand2 < 0.5 \\ best\left(x_j\right) \times (MoPr) \times \left(\left(U_j - L_j\right) \times \mu + L_j\right), Otherwise \end{cases} \tag{4.9}$$

whereas $x_i(t+1)$ denotes the i^{th} solution of $(t+1)^{th}$ iteration, $x_{i,j}(t)$ stands for the j^{th} location of i^{th} individual at the current generation. $best(x_j)$ implies the j^{th} position of finest solution. ϵ indicates the small positive integer, the upper and lower boundaries of j^{th} location is defined by U_j and L_j, correspondingly [180]. μ that is, the controlling parameter was fixed to 0.5. The Math Optimizer Probability ($MoPr$), which is a co-efficient, is evaluated as:

$$MoPr(t) = 1 - \frac{t^{\frac{1}{\theta}}}{M\_Iter^{\frac{1}{\theta}}} \quad (4.10)$$

In which, $MoPr(t)$ represents the value of $MoPr$ function at t^{th} iterations. M_Iter signifies the maximal iteration count [181]. θ refers to the vital parameter which manages the exploitation efficacy throughout iterations.

In AOA, utilizing Subtraction (S) or Addition (A) operators, the exploitation approach was established as expressed in Eq. (11). At this point, μ is also a constant that is set equivalent to 0.5.

$$x_{i,j}(t+1) = \begin{cases} best(x_j) - (MoPr) \times \left((U_j - L_j) \times \mu + L_j \right), & rand3 < 0.5 \\ best(x_j) - (MoPr) \times \left((U_j - L_j) \times \mu + L_j \right), & Otherwise \end{cases} \quad (4.11)$$

The AOA technique was defined in Algorithm 1, dependent upon the presented exploration and exploitation approaches [182].

Algorithm 1: Procedure of Arithmetic Optimization Algorithm (AOA)

Initializing parameters μ and θ of the algorithm. Take $t = 0$.
Initializing the n count of solutions positions arbitrarily as per Eq. (7).
While (end condition could not be met) do
  Evaluate the fitness values of created solutions.
  Save the optimum solution accomplished so far.
  Eq. (4.8) is utilized for modifying the $MoAc$ value.
  Eq. (4.10) is utilized for modifying the $MoPr$ value.
  For ($i = l$ to Solutions) do
    For ($J = 1$ to Positions) do
    Create an arbitrary value ($i.e. randl, rand2, and\ rand3$).
    if ($randl > MoAc$)
    //Exploration stage
      if(rand2 > 0.5)
      (1) Employ the Division operator (D " ÷ ").
        Utilizing the 1st rule of Eq. (4.9), upgrade the i^{th} solution positions.
      else
      (2) Employ the Multiplication operator (M " × ").
        Utilizing the 2nd rule of Eq. (4.9) $4^{\mathcal{M}}$ upgrade the i^{th} solution positions.

End if

Else

//Exploitation stage

if ($r3 > 0.5$)

(1) Employ the Subtraction operator ($S" - "$).

Utilizing the 1st rule of Eq. (4.11), upgrade the i^{th} solution positions.

Else

(2) Employ the Addition operator ($A'' + ''$).

Utilizing the 2nd rule of Eq. (4.11), upgrade the i^{th} solution positions.

End if

End if

End for

End for

$r = r + 1$

end While

Return the global optimum solution.

4.4 Results and Discussion

Two datasets are used to assess the experimental validity of the ASDS-AOADBN approach: DIHARD dataset [183] and CALLHOME dataset [184].

The Speaker Recognition Database (DIHARD) contains single channel input level improvement; likewise, estimation sets comprise 5-10 mins duration instances obtained in 11 conversational domains, each containing around two hours of audio. The National Institute of Standards and Technology (NIST) runs an evaluation series for diarization systems called the DIHARD (Diarization Hard) challenge series. To evaluate the resilience of diarization systems, DIHARD datasets usually contain a range of acoustic situations, including disparate settings, overlapping speech, and difficult scenarios.

The CALLHOME database is a group of multi-lingual telephone information that takes 500 recordings; the range of all the recordings is 2-5 mins per file. The speaker counts in all the recording differs from 2-7, with maximum files having two speakers.

Table 1 displays the overall tracking distance efficiency (TDE) results of the ASDS-AOADBN technique with recent models [185]. Figure 4.5 represents the comparative TDE results of the ASDS-AOADBN technique on the DIHARD dataset. The results indicate that the SDS-TDNN model reaches poor results with increasing TDE results, while the Speaker diarization system

with modified conditional generative adversarial *network* (SDS-MCGAN) model attains slightly reduced TDE values. Although the Speaker diarization system utilizing Holo-entropy with the extended Linear Prediction with deep neural network (SDS-HXLP-DNN) and Speaker diarization system utilizing Holo-entropy with the extended Linear Prediction with deep convolutional neural networks using sailfish optimizer Algorithm (SDS-HXLP-DCNN-SOA) approaches accomplish closer results, the ASDS-AOADBN technique shows effective results with the least TDE values. For instance, with 0.03 frame length, the ASDS-AOADBN technique offers a lower TDE of 1200 while the Speaker diarization system with time-delay neural network (SDS-TDNN) [186], SDS-MCGAN [187], SDS-HXLP-DNN [188] and SDS-HXLP-DCNN-SOA [189] models obtain increasing TDE of 8100, 5500, 2200, and 1600 respectively.

4.4.1 Performance metrics

The automated speaker identification architecture's effectiveness was evaluated using essential metrics, including precision, accuracy, diarization error rate (DER), Jaccard error rate (JER), Tracking Distance Efficiency (TDS), False Alarm Rate Efficiency (FARE) and Diarization Error Rate Efficiency (DERE).

TDS: The Automated Speaker Diarization System's Tracking Distance Efficiency quantifies the system's ability to precisely track and distinguish between speakers over an extended period of time. A statistic that combines the tracking distance and the accurate speaker assignment is frequently used to quantify performance.

$$TDE = \frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{C_i} \tag{4.12}$$

N is the total number of speakers, D_i is the total tracking distance for speaker i, and C_i is the total number of correctly assigned speakers for speaker i.

FARE: In an Automated Speaker Diarization System, the term FARE describes how well the system minimizes false alarms or inaccurate speaker change detections. Typically, the following equation is used to measure it:

$$FARE = \frac{False\ \ Alarms}{Total\ \ Detected\ \ Speaker\ \ Changes} \tag{4.13}$$

DERE: In an audio recording, DERE is a metric used to quantify how accurately a system segments and labels speakers. The calculation involves expressing the percentage of accurately identified speaker segments to the whole speech time. For DER Efficiency, the formula is:

$$DERE = \frac{Total\ \ Speech\ \ Duration - DER}{Total\ \ Speech\ \ Duration} \times 100 \tag{4.14}$$

Precision: It is a measurement of the percentage of unfavourable outcomes that were expected. In the majority of instances, the precision measure is used in order to quantify exactness. In proportion to the drop in the FP rate, the accuracy values rise.

$$\Pr esision = \frac{TP}{FP + TP} \tag{4.15}$$

Accuracy: The accuracy was determined by dividing the total number of samples by the sum of true positives (TP) and true negatives (TN).

$$Accuracy = \frac{(TP + TN)}{(TN + TP + FN + FP)} \tag{4.16}$$

Diarization Error Rate: DER, which is the total of three distinct error kinds, is used to assess the correctness of speaker diarization systems: Speech-related false alarms (FA), missing speech detection, and misidentification of speaker label

$$DER = \frac{FA + Missed + Spea\,\mathrm{ker} - Confusion}{Total\ \ Duration\ of\ \ Time} \tag{4.17}$$

Jaccard Error Rate: The DIHARD II assessment introduced the JER metric for the first time. JER strives to provide equal weighting for each speaker throughout the evaluation process. Per-speaker error rates are calculated individually for each speaker and then averaged to get the Joint Error Rate (JER), as opposed to the Diarization Error Rate (DER), which is calculated for the whole utterance.

$$JER = \frac{1}{N} \sum_{i}^{N_{ref}} \frac{FA_i + MISS_i}{TOTAL_i} \tag{4.18}$$

$TOTAL_i$ in Eq.(4.18) is the sum of the speaking time of the i-th speaker in the hypotheses and the i-th speaker in the reference transcript. The reference script has a total of N_{ref} speakers. Keep in mind that the part of FA_i used in the JER computation reflects the Speaker-Confusion in DER. JER never goes over 100%, however DER can grow well above 100% since it uses a union operation between the reference and the hypotheses. The supplied audio recording has a strong correlation between DER and JER; however, in the event that a subgroup of speakers dominates, JER naturally rises above the average.

Table 4.1 TDE analysis of ASDS-AOADBN approach with other systems under two datasets

Tracking Distance Efficiency					
Frame length	**SDS-TDNN**	**SDS-MCGAN**	**SDS-HXLP-DNN**	**SDS-HXLP-DCNN-SOA**	**ASDS-AOADBN**
DIHARD Dataset					
0.03	8100	5500	2200	1600	1200
0.06	8100	5900	2300	1900	1500
0.09	8300	6100	2900	2600	2400
0.12	8500	6900	4200	2600	2500
0.15	8500	8000	4400	3000	2700
CALLHOME Dataset					
0.03	6400	7800	6500	3800	3300
0.06	7800	8300	7600	4300	3500
0.09	8100	11800	8700	5000	4100
0.12	8700	12100	10500	5800	5400
0.15	10700	15800	12500	6000	5500

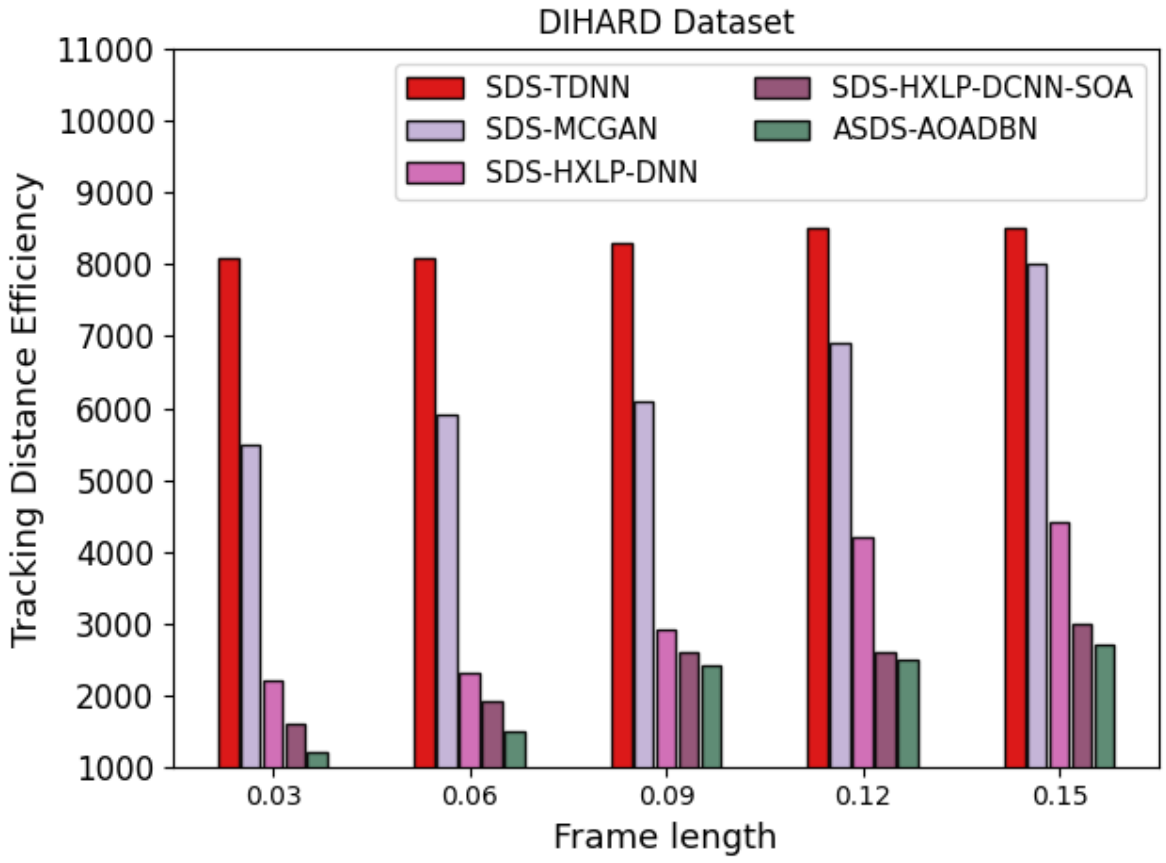

Figure 4.5 TDE analysis of ASDS-AOADBN approach under DIHARD dataset

Figure 4.6 compares TDE outcomes of the ASDS-AOADBN algorithm on the CALLHOME dataset. The outcomes state that the SDS-TDNN algorithm gains worse effects with enhancing TDE results, while the SDS-MCGAN approach attains slightly reduced TDE values. Although the SDS-HXLP-DNN and SDS-HXLP-DCNN-SOA techniques accomplish closer results, the ASDS-AOADBN system shows effective results with the least TDE values. For instance, with 0.03 frame length, the ASDS-AOADBN technique offers a lesser TDE of 3300, while the SDS-TDNN, SDS-MCGAN, SDS-HXLP-DNN, techniques obtain maximal TDE of 6400, 7800 and 6500, correspondingly.

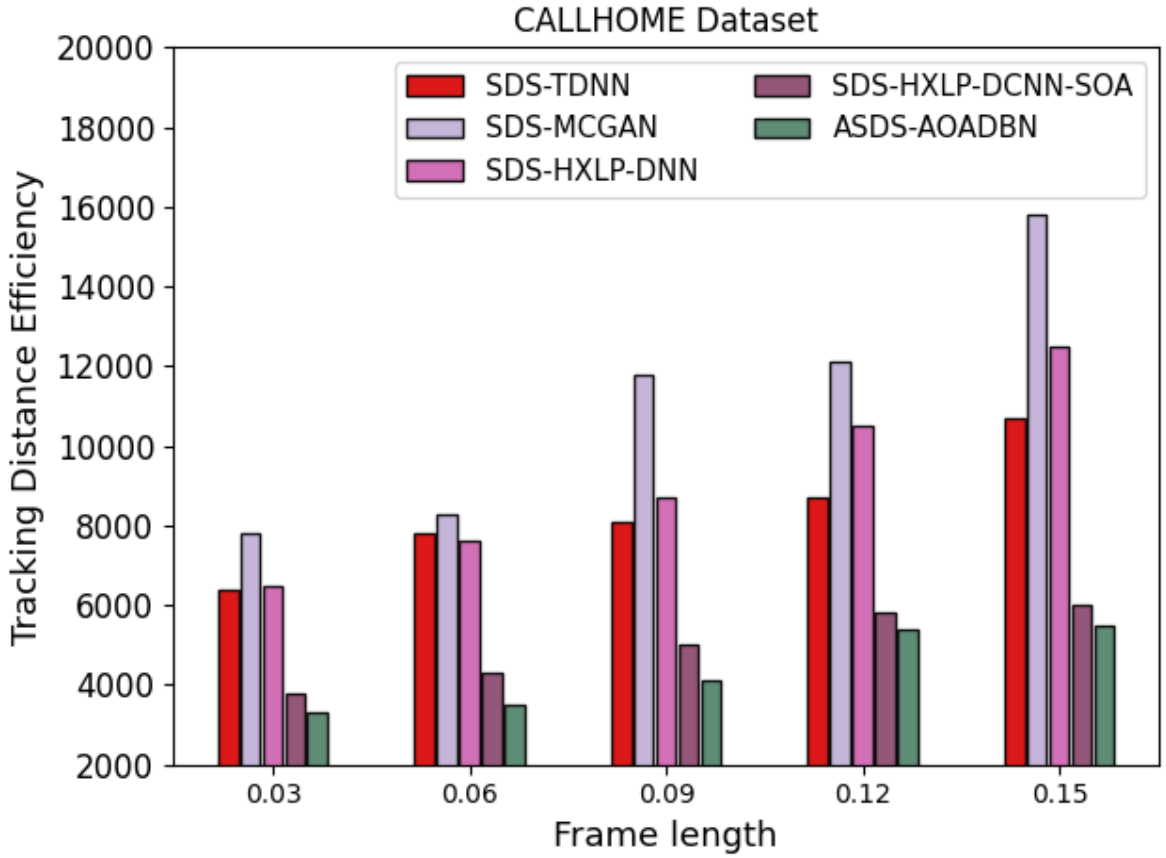

Figure 4.6 TDE analysis of ASDS-AOADBN approach under CALLHOME dataset

Table 4.2 demonstrates the overall FARE outcomes of the ASDS-AOADBN system with recent approaches. Figure 4.7 signifies the comparative FARE results of the ASDS-AOADBN technique on the DIHARD dataset. The results indicate that the SDS-TDNN model reaches poor results with increasing FARE results, while the SDS-MCGAN model attains slightly reduced FARE values. Although the SDS-HXLP-DNN and SDS-HXLP-DCNN-SOA methods accomplish closer results, the ASDS-AOADBN technique shows effective results with the least FARE values. For instance, with a 5.50 frame length, the ASDS-AOADBN technique offers a

lower FARE of 0.103, while the SDS-TDNN, SDS-MCGAN, SDS-HXLP-DNN, models obtain increasing FARE of 0.465, 0.406 and 0.295, correspondingly.

Table 4.2 FARE analysis of ASDS-AOADBN approach with other systems under two datasets

False Alarm Rate Efficiency					
Frame length	**SDS-TDNN**	**SDS-MCGAN**	**SDS-HXLP-DNN**	**SDS-HXLP-DCNN-SOA**	**ASDS-AOADBN**
DIHARD Dataset					
5.50	0.465	0.406	0.295	0.287	0.103
7.00	0.497	0.443	0.375	0.333	0.184
8.50	0.594	0.476	0.354	0.342	0.217
10.00	0.611	0.491	0.436	0.371	0.292
11.50	0.652	0.522	0.489	0.390	0.294
CALLHOME Dataset					
5.50	0.552	0.474	0.423	0.240	0.154
7.00	0.563	0.487	0.457	0.248	0.171
8.50	0.573	0.490	0.485	0.270	0.189
10.00	0.589	0.531	0.512	0.295	0.249
11.50	0.610	0.567	0.523	0.358	0.339

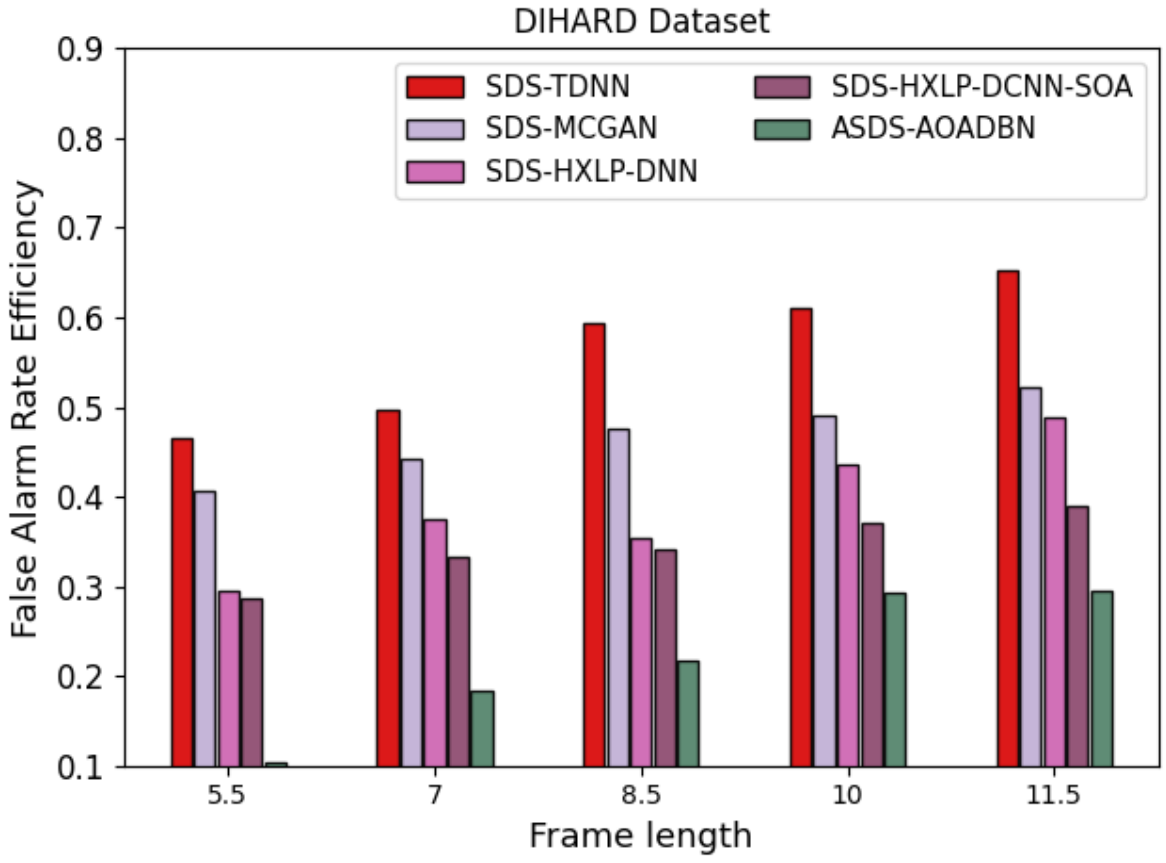

Figure 4.7 FARE analysis of ASDS-AOADBN approach under DIHARD dataset

Figure 4.8 exemplifies the comparative FARE outcomes of the ASDS-AOADBN technique on the CALLHOME dataset. The outcomes inferred that the SDS-TDNN algorithm achieves worse outcomes with maximal FARE results while the SDS-MCGAN approach attains somewhat lesser FARE values. The SDS-HXLP-DNN and SDS-HXLP-DCNN-SOA algorithms realize closer results, and the ASDS-AOADBN technique outperformed effectual outcomes with the least FARE values. For instance, with a 5.50 frame length, the ASDS-AOADBN approach offers a minimal FARE of 0.154, while the SDS-TDNN, SDS-MCGAN, SDS-HXLP-DNN, techniques reach a maximal FARE of 0.552, 0.474 and 0.423, correspondingly.

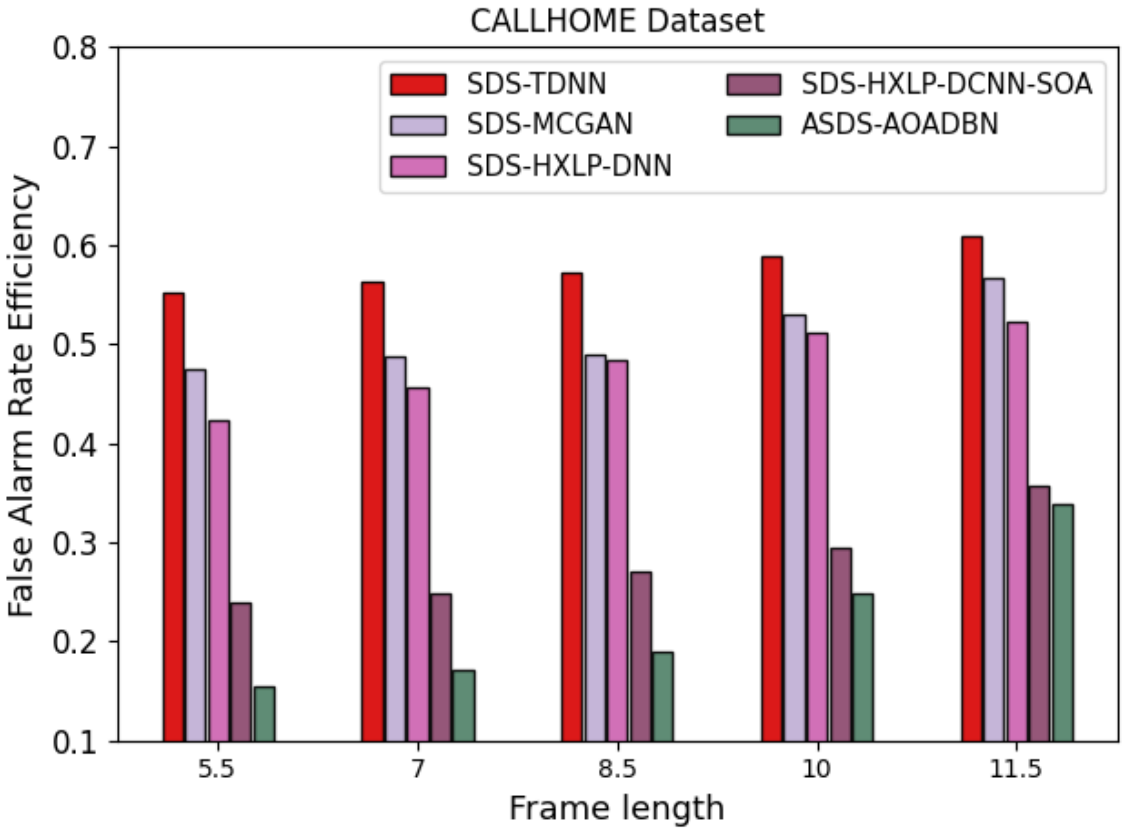

Figure 4.8 FARE analysis of ASDS-AOADBN approach under CALLHOME dataset

Table 4.3 gives the overall diarization error rate efficiency (DERE) outcomes of the ASDS-AOADBN system with recent techniques. Figure 4.9 demonstrates the comparative DERE outcomes of the ASDS-AOADBN technique on the DIHARD dataset. The results implied that the SDS-TDNN system reaches worse outcomes with increasing DERE results while the SDS-MCGAN methodology attains slightly reduced DERE values. Moreover, the SDS-HXLP-DNN and SDS-HXLP-DCNN-SOA techniques accomplish closer results, and the ASDS-AOADBN technique depicts effectual outcomes with minimal DERE values. For instance, with a 5.50

frame length, the ASDS-AOADBN technique offers lower DERE of 0.103, while the SDS-TDNN, SDS-MCGAN, SDS-HXLP-DNN, techniques obtain maximal DERE of 0.465, 0.406 and 0.295, correspondingly.

Table 4.3 DERE analysis of the ASDS-AOADBN method with other systems

Diarization Error Rate Efficiency					
Frame length	**SDS-TDNN**	**SDS-MCGAN**	**SDS-HXLP-DNN**	**SDS-HXLP-DCNN-SOA**	**ASDS-AOADBN**
DIHARD Dataset					
6	0.605	0.429	0.678	0.253	0.221
8	0.608	0.451	0.682	0.254	0.229
10	0.638	0.496	0.751	0.267	0.255
12	0.640	0.547	0.774	0.291	0.267
14	0.729	0.595	0.794	0.301	0.290
CALLHOME Dataset					
6	0.608	0.403	0.603	0.214	0.183
8	0.616	0.417	0.730	0.230	0.216
10	0.681	0.461	0.755	0.247	0.225
12	0.705	0.552	0.874	0.324	0.304
14	0.725	0.553	0.922	0.370	0.339

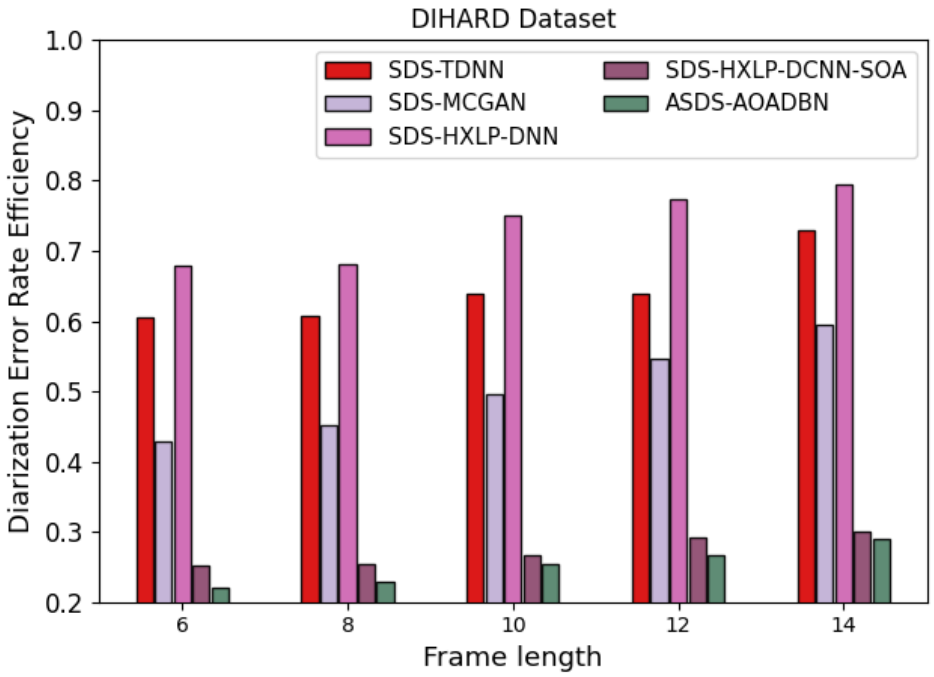

Figure 4.9 DERE analysis of ASDS-AOADBN approach under the DIHARD dataset

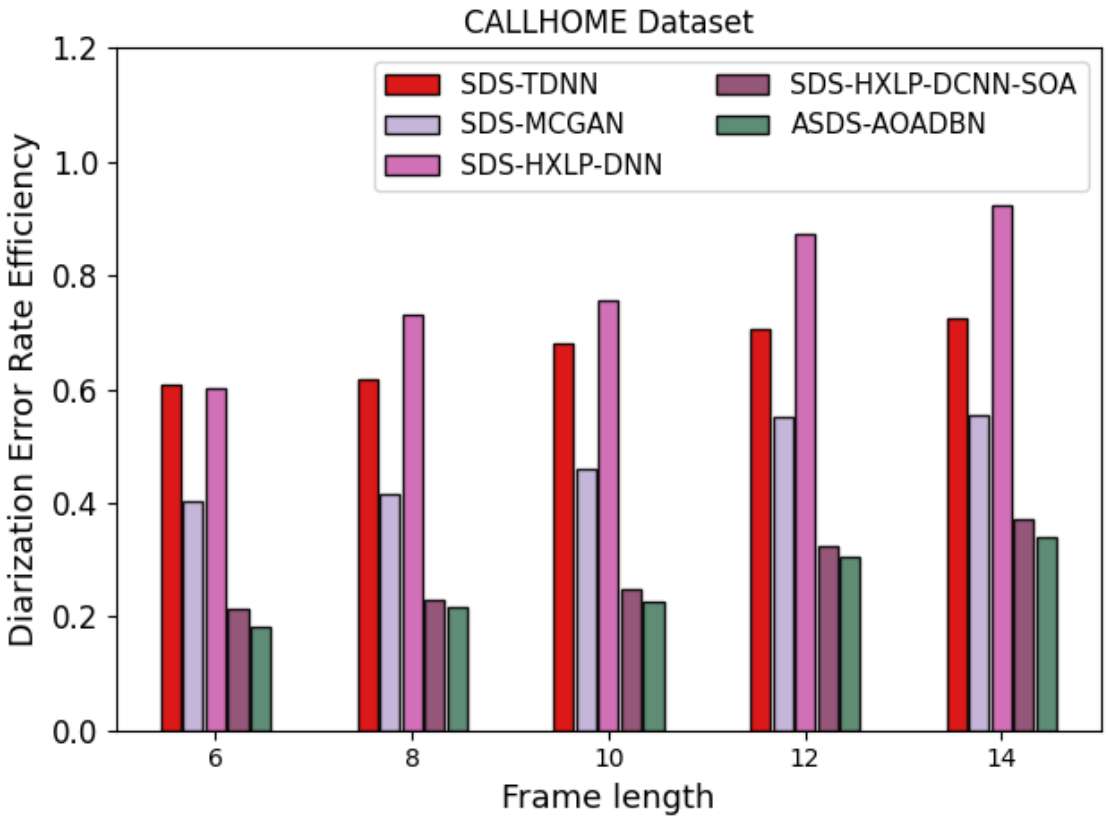

Figure 4.10 DERE analysis of ASDS-AOADBN approach under CALLHOME dataset

Figure 4.10 depicts the comparative DERE results of the ASDS-AOADBN technique on the CALLHOME dataset. The outcomes inferred that the SDS-TDNN technique reaches worse outcomes with enhancing DERE outcomes while the SDS-MCGAN system attains somewhat reduced DERE values. Furthermore, the SDS-HXLP-DNN and SDS-HXLP-DCNN-SOA approaches realize closer results, and the ASDS-AOADBN technique exhibits effectual outcomes with minimum DERE values. For instance, with a 5.50 frame length, the ASDS-AOADBN method offers lower DERE of 0.154, while the SDS-TDNN, SDS-MCGAN, SDS-HXLP-DNN, techniques acquire increasing DERE of 0.552, 0.474 and 0.423, correspondingly. Table 4.4 and Figure 4.11 signify the comparative diarization error rate (DER) outcomes of the ASDS-AOADBN system with existing. The outcomes represent that the SDS-TDNN technique gains worse outcomes with superior DER results while the SDS-MCGAN algorithm reaches somewhat lesser DER values. But, the SDS-HXLP-DNN and SDS-HXLP-DCNN-SOA approaches realize closer results, and the ASDS-AOADBN technique demonstrates efficacy results with the least DER values. For instance, with speaker 1, the ASDS-AOADBN technique offers a lower DER of 0.353, while the SDS-TDNN, SDS-MCGAN and SDS-HXLP-DNN approaches obtain increasing DER of 6.329, 2.16, 1.622 correspondingly.

Table 4.4 DER analysis of ASDS-AOADBN approach with other systems

Diarization Error Rate vs. Count of Speakers					
No. of Speakers	**SDS-TDNN**	**SDS-MCGAN**	**SDS-HXLP-DNN**	**SDS-HXLP-DCNN-SOA**	**ASDS-AOADBN**
Speaker 1	6.329	2.16	1.622	0.549	0.353
Speaker 2	3.871	1.307	2.898	1.304	0.951
Speaker 3	8.399	1.594	2.193	1.518	1.079
Speaker 4	8.165	9.578	3.432	1.385	1.222
Speaker 5	4.489	7.212	5.771	3.835	3.242

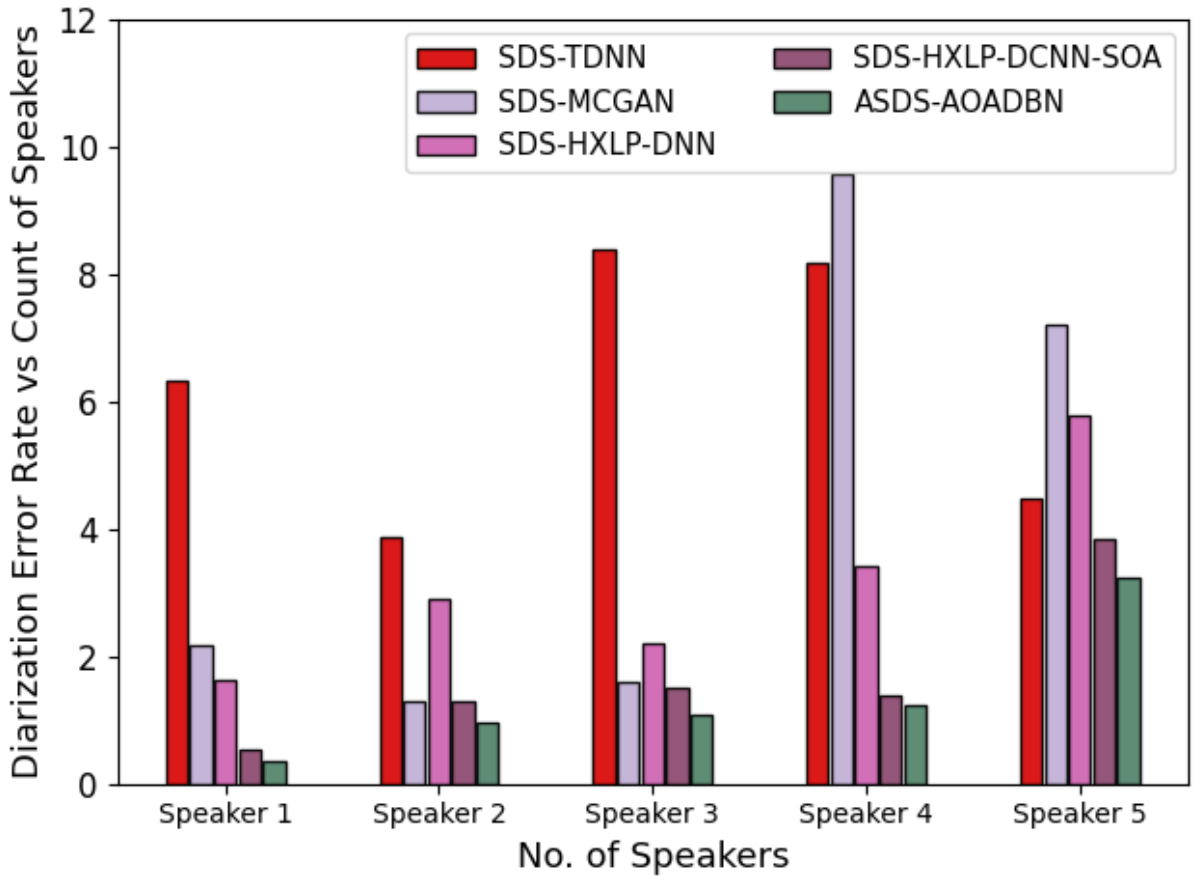

Figure 4.11 DER analysis of ASDS-AOADBN approach under varying speakers

In Table 4.5, the results of the ASDS-AOADBN technique are compared with other models in terms of TP, FP, and FN. The experimental values indicate that the ASDS-AOADBN technique performs better with TP of 0.963, FP of 0.212, and FN of 0.189.

Table 4.5 Classifier outcome of ASDS-AOADBN approach with other techniques

Methods	True Positive	False Positive	False Negative
SDS-TDNN	0.922	0.798	0.669
SDS-MCGAN	0.957	0.779	0.379
SDS-HXLP-DNN	0.723	0.508	0.751
SDS-HXLP-DCNN-SOA	0.822	0.250	0.270
ASDS-AOADBN	0.963	0.212	0.189

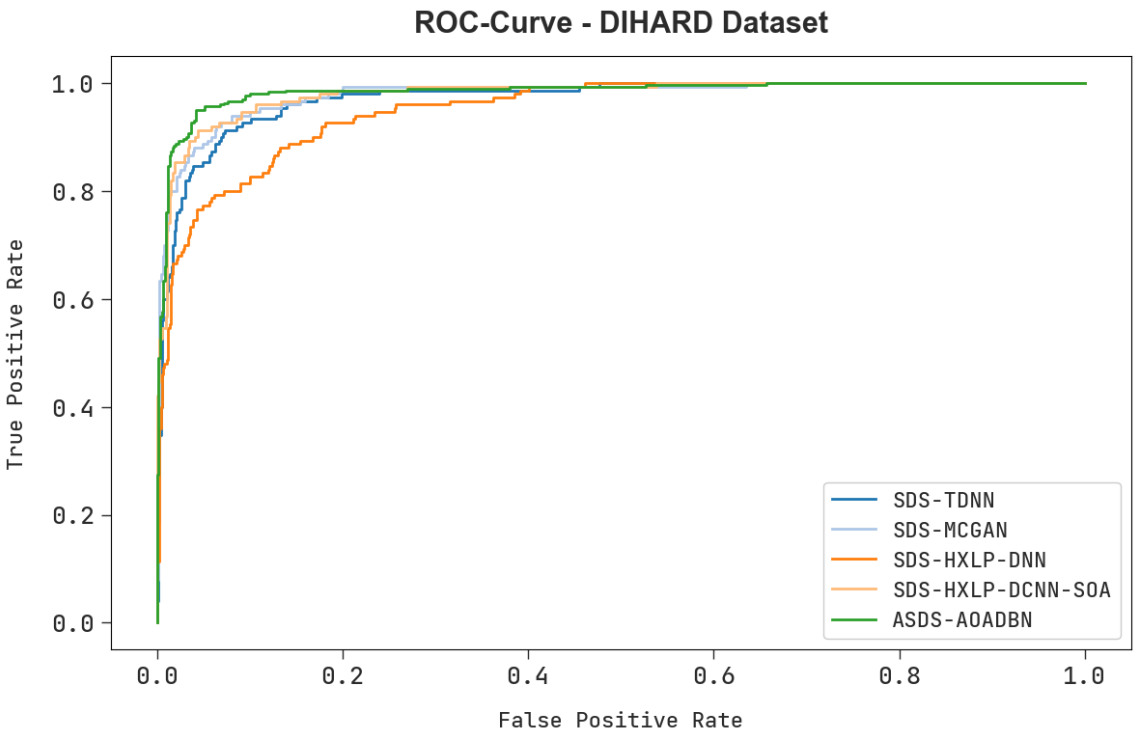

Figure 4.12 ROC curve of ASDS-AOADBN approach under DIHARD dataset

The ASDS-AOADBN method's ROC analysis on the DIHARD dataset is shown in Figure 4.12. The improved ROC values were defined by the figure as a consequence of the ASDS-AOADBN system. On top of that, the ASDS-AOADBN method may increase ROC values across the board.

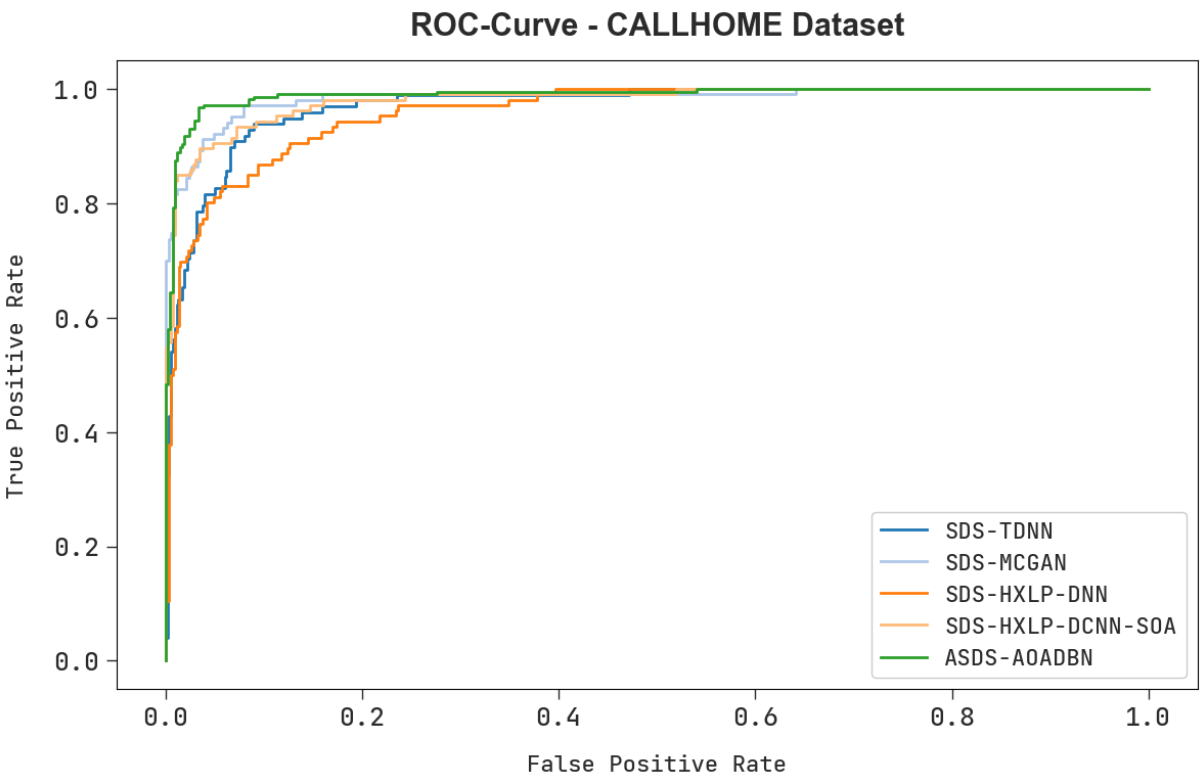

Figure 4.13 ROC curve of ASDS-AOADBN approach under CALLHOME dataset

On the CALLHOME dataset, the ROC findings of the ASDS-AOADBN method are shown in Figure 4.13. It was implied from the figure that the ASDS-AOADBN system produced higher ROC values. Not to mention that the ASDS-AOADBN approach clearly extends better ROC values across the board. These results confirmed that the ASDS-AOADBN method outperformed more contemporary techniques across a variety of metrics.

Table 4.6 Comparison Analysis of precision and accuracy analysis for ASDS-AOADBN method

Methods	**DIHARD Dataset**		**CALLHOME Dataset**	
	Precision	**Accuracy**	**Precision**	**Accuracy**
SDS-TDNN	0.85	0.88	0.87	0.89
SDS-MCGAN	0.88	0.90	0.90	0.92
SDS-HXLP-DNN	0.87	0.80	0.89	0.91
SDS-HXLP-DCNN-SOA	0.89	0.91	0.91	0.93
ASDS-AOADBN	0.91	0.93	0.92	0.94.

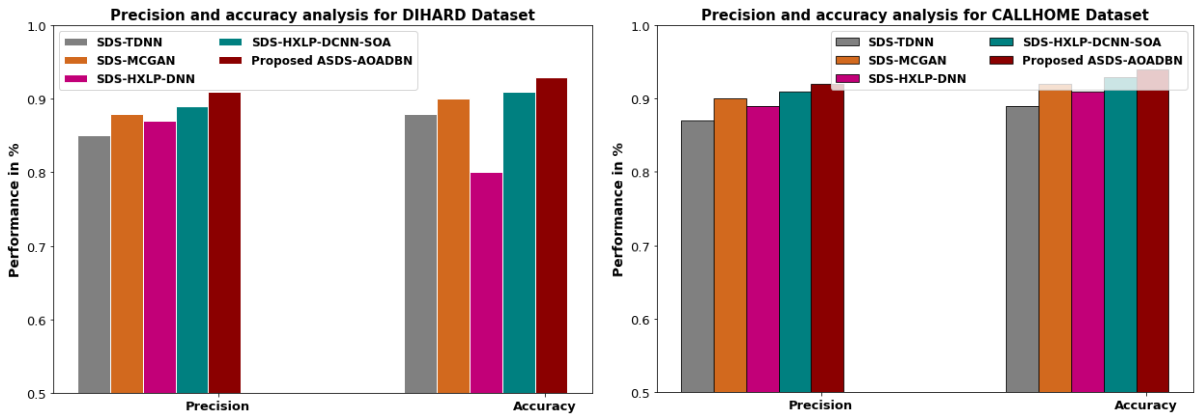

Figure 4.14 Comparison Analysis of precision and accuracy analysis for ASDS-AOADBN method

In Figure 4.14 and Table 4.6, the ASDS-AOADBN technique's precision and accuracy are contrasted with other widely used methods. The graph illustrates how the DL approach has an enhanced performance with precision and accuracy. For instance, the ASDS-AOADBN model's precision value for DIHARD dataset data is 0.91%, whereas the precision values for the SDS-TDNN, SDS-HXLP-DNN and SDS-HXLP-DCNN-SOA models are 0.85%, 0.87% and 0.89%. Similarly, the ASDS-AOADBN model's accuracy value for DIHARD dataset data is 0.93%, whereas the accuracy values for the SDS-TDNN, SDS-HXLP-DNN and SDS-HXLP-DCNN-SOA models are 0.88%, 0.80%, and 0.91%.

For instance, the ASDS-AOADBN model's precision value for CALLHOME dataset data is 0.92%, whereas the precision values for the SDS-TDNN, SDS-HXLP-DNN and SDS-HXLP-DCNN-SOA models are 0.87%, 0.89% and 0.91%. Similarly, the ASDS-AOADBN model's accuracy value for CALLHOME dataset data is 0.94%, whereas the accuracy values for the SDS-TDNN, SDS-HXLP-DNN and SDS-HXLP-DCNN-SOA models are 0.89%, 0.91% and 0.93%.

Table 4.7 Comparison Analysis of DER and JER analysis for ASDS-AOADBN method

Methods	**DIHARD Dataset**		**CALLHOME Dataset**	
	DER	**JER**	**DER**	**JER**
SDS-TDNN	0.75	0.67	0.71	0.60
SDS-MCGAN	0.80	0.65	0.69	0.56
SDS-HXLP-DNN	0.74	0.66	0.65	0.52
SDS-HXLP-DCNN-SOA	0.72	0.68	0.62	0.54
ASDS-AOADBN	0.65	0.60	0.60	0.51

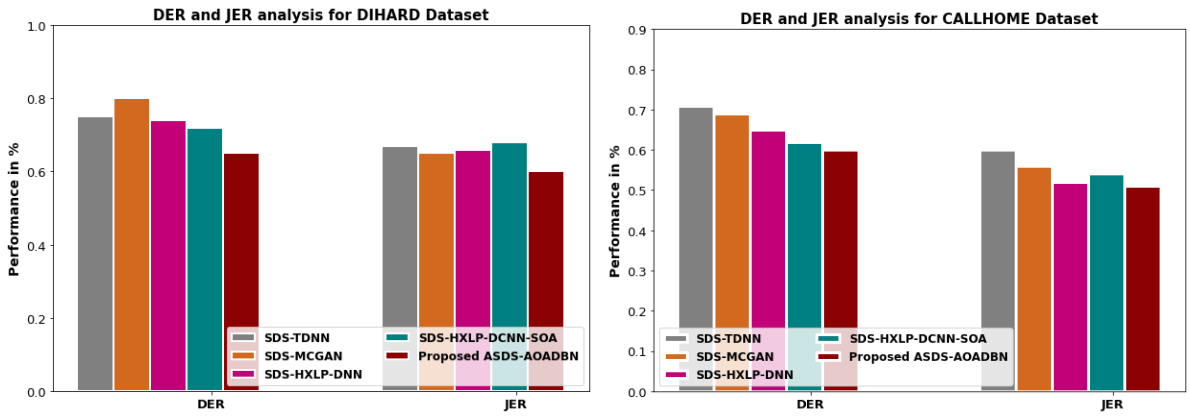

Figure 4.15 Comparison Analysis of DER and JER analysis for ASDS-AOADBN method

Figure 4.15 and Table 7 display an DER and JER comparison of the ASDS-AOADBN strategy with other well-known methods. The DL technique has an enhanced performance while reducing DER and JER, as shown in the graph. For instance, the ASDS-AOADBN model's DER value for DIHARD dataset data is 0.65%, while the DER values for the SDS-TDNN, SDS-MCGAN and SDS-HXLP-DNN and models are 0.75%, 0.80% and0.74%, respectively. The ASDS-AOADBN model, however, has demonstrated its best performance for various data sizes with low DER and JER values. In a similar vein, for DIHARD dataset data, the DER value for the ASDS-AOADBN is 0.60%, whereas, for the SDS-TDNN, SDS-MCGAN, and SDS-HXLP-DNN models, it is 0.67%, 0.65% and 0.66%, respectively.

For instance, the ASDS-AOADBN model's DER value for CALLHOME dataset data is 0.60%, while the DER values for the SDS-TDNN, SDS-MCGAN and SDS-HXLP-DNN models are 0.71%, 0.69% and 0.65%, respectively. The ASDS-AOADBN model, however, has demonstrated its best performance for various data sizes with low DER and JER values. In a similar vein, for CALLHOME dataset data, the DER value for the ASDS-AOADBN is 0.51%, whereas, for the SDS-TDNN, SDS-MCGAN and SDS-HXLP-DNN models, it is 0.60%, 0.56% and 0.52%, respectively.

4.5 Discussion and Applications

The proposed Automatic Speaker Diarization System demonstrates promising results in improving the accuracy and effectiveness of speaker diarization procedures. It does this by relating a DBN model with an Arithmetic Optimization Algorithm. By incorporating the AOA, the system efficiently refines the audio segment clustering, reducing the possibility of errors in speaker identification. The DBN model further increases system adaptability by removing intricate patterns and features from the audio data. Associating the experimental findings with

conventional methods, it is clear that the precision of diarization has expressively improved. Speaker overlap and variable acoustic conditions are among the problems that the DBN and AOA are designed to solve. Because of the system's scalability, it may be applied to a multitude of tasks, including surveillance, meeting transcription, and voice recognition. The proposed method has a great deal of potential to develop automated speaker diarization technology and offer a dependable, practical solution for practical applications, as per the research findings. The experimental results show an accuracy of 0.94 for the CALLHOME dataset and 0.93 for the DIHARD dataset.

4.6 Summary

In this chapter, a novel ASDS-AOADBN system for the automatic speaker diarizaion process, which accurately identifies and classifies the speaker signal from the input audio signals is established. To accomplish this, the presented ASDS-AOADBN technique primarily preprocessed the audio input signals using the SGF technique, eradicating the noise and improving the audio input quality. Followed by the MFCC features are derived from the preprocessed ones, which are then passed into the DBN model for the speaker identification process. Finally, the AOA-based hyperparameter tuning process is performed to adjust the hyperparameter values of the DBN approach. The experimental result analysis stated the better performance of the ASDS-AOADBN technique over current state-of-the-art DL models. The comparative analysis showed better performance of the ASDS-AOADBN technique over the current state-of-the-art DL models. Simplifying the extraction of complex audio elements has allowed DL to improve the system's speaker recognition performance. With its ability to adapt and learn from different datasets, the model is flexible in real-world circumstances where speech appearances vary. The proposed technology has been demonstrated experimentally to outperform existing methods; potential applications include voice assistants, security systems, and transcribing services. Scalability, flexibility, and effectiveness of the system are critical in the dynamic field of automated speaker diarization. In the future, the ASDS-AOADBN technique's performance will be boosted by employing segmentation approaches.

5. Chapter - Conclusion and Scope of Future Work

5.1 Conclusion

The prime goal of this research is to create a reliable, accurate, and robust speaker recognition system that can be efficiently used in forensic applications. The other goal is to investigate the many components of the speaker recognition system that have an impact on the system's overall performance. The following final thoughts emerge from a review of the thesis work:

The First Chapter offers a comprehensive analysis of several biometric and forensic speaker recognition methods. The text provides a thorough examination of several important elements, such as speech acoustics, mathematical and statistical methods, and approaches developed for speaker recognition. Furthermore, it examines methods for modelling speakers and investigates the use of deep neural networks in the field of speaker recognition. In addition, the chapter carries out a comprehensive review of the existing literature on speaker identification, verification, and diarization. The purpose of this study is to have a detailed knowledge of the many difficulties that exist in modern speaker identification models. The chapter aims to clarify the intricacies of effectively recognising, confirming, and segmenting speakers by using current research and approaches to analyse their distinct speech features.

The chapter not only offers an in-depth exploration of biometric and forensic speaker recognition systems but also sheds light on their diverse applications. It elucidates how speaker recognition systems find utility across various domains, including but not limited to security access control, forensic investigations, telecommunication systems, and human-computer interaction. By highlighting these practical applications, the chapter underscores the relevance and significance of speaker recognition technology in addressing real-world challenges and enhancing the efficiency of numerous systems and processes.

In essence, the First Chapter establishes a fundamental structure for research and studies in the area of speaker recognition. This article provides a foundation for understanding the theoretical foundations, methodological methods, and practical difficulties related to creating and using speaker recognition systems in both biometric and forensic settings.

The second chapter explains a Speaker Identification framework utilizing a Mask R-CNN classifier optimized through Hosted Cuckoo Optimization (HCO). This approach aims to enhance accuracy and signal robustness in speaker identification tasks. Initially, input speech signals are extracted from a real-time speaker recognition database. To bolster signal robustness, four types of features, namely MFDPCC, GFCC, PNCC, and Spectral entropy, are

extracted from the input speech signals. Subsequently, the Mask R-CNN classifier is employed for speaker ID classification, with the HCO algorithm optimizing its parameters.

The methodology is then applied in real-time applications such as telephone banking and fax mailing, with simulations conducted using MATLAB. The Mask-R-CNN-HCO technique exhibits favourable outcomes, surpassing recent methods such as KNN, MCSVM, GMMCNN, DNN, and GMMDNN in terms of Recall, Specificity, and F-Score for automated classification of speaker identification. Specifically, the proposed method outperforms existing methods by notable margins, highlighting its efficacy in speaker identification tasks.

The third chapter of the document elaborates on the TTFEM-AFSV model, designed specifically for speaker verification in forensic contexts. The model first applies the AMF approach to remove background noise from audio sources. Next, it feeds the spectrograms and Mel-frequency cepstral coefficients (MFCC) into the Inception v3 model, a CNN. By using the ALO technique, the Inception v3 model's hyper-parameters are fine-tuned. Once this is done, the model uses an LSTM-RNN classifier for automatic speaker verification. The efficacy of the TTFEM-AFSV model is evaluated through a series of experiments. Comparative analysis indicates a significant improvement in performance compared to recent approaches. Consequently, the TTFEM-AFSV model proves to be highly effective for real-time ASR in forensic applications.

The ASDS-AOADBN system is described in detail in *the fourth chapter*. This system is an automatic speaker diarization tool that properly differentiates and categorises speaker signals within the audio data that is supplied. The ASDS-AOADBN methodology starts with the preprocessing of audio inputs using the SGF method. This successful removal of noise and enhancement of overall audio quality is the first step in the procedure. After that, MFCC characteristics are retrieved from the signals that have been preprocessed and then put into a Deep Belief Network (DBN) model for the purpose of speaker identification. An AOA-based tuning technique is used in order to achieve the process of fine-tuning the hyperparameters of the DBN model. According to the findings of the experiments, the ASDS-AOADBN approach demonstrates higher performance when compared to the Deep Learning (DL) models that are currently considered to be state-of-the-art. Through the simplification of the extraction of complex audio aspects, DL has been able to improve speaker identification inside the system. Because of its versatility and its ability to learn from a wide variety of datasets, the model is well suited for situations that occur in the real world and include speech features that fluctuate. The proposed technology has shown its superiority over the approaches that are now in use via

the process of experimental validation. It has the potential to be used in a variety of applications, including voice assistants, security systems, and transcription services.

In the constantly shifting world of automated speaker diarization, scalability, adaptability, and effectiveness are of the utmost importance, and the ASDS-AOADBN system is specifically designed to meet these needs in an efficient manner.

5.2 Scope of future work

The research presented in this thesis advances the field of speaker recognition systems, especially with regard to improving the efficiency of various speaker recognition models. This is achieved by strengthening and refining a variety of existing speech recognition models. Listed below are some potential areas of future study:

a. It is well recognised that voiced parts of speech include more distinguishable information about the speaker's identity in comparison to unvoiced segments. Obtaining accurate speech samples may be difficult in many cases, especially when considering the characteristics of speech. This feature has not been extensively examined in the research. However, it is suggested that future efforts might use unvoiced sound segments for the purpose of identifying speakers in forensic applications. These portions include a range of human noises, such as weeping, hiccuping, yawning, snoring, sighing, burping, gasping, grunting, groaning, and whimpering. Each of these sounds has unique acoustic properties that may be efficiently used for the goal of speaker identification. In addition, using unvoiced sound segments broadens the range of information available for analysis, hence improving the precision and dependability of speaker identification approaches.
b. For the goal of speaker diarization, another area that needs to be addressed is the issues that are related to appropriately segmenting audio recordings into segments that are unique to the speaker. On the subject of achieving precise segmentation, there are a number of aspects that provide a significant amount of difficulty. It is possible to experience overlapping speech, background noise, and changes in the features of one's speech.
c. In order to improve the TTFEM-AFSV model's overall performance, future improvements to speaker verification might include creating an ensemble of fusion-based Deep Learning (DL) models.

d. The future trajectory of the ongoing research in speaker recognition encompasses many critical goals. Creating an immense database that has a substantial number of speakers is one of the alternatives that may be considered. The database will be carefully organised to include voice data obtained from different sensors in diverse regions strategically placed at varying distances from each other. By using this method, it is expected to include a wide variety of sound situations and circumstances, therefore improving the strength and applicability of the database. Moreover, goal is to establish an all-encompassing multilingual repository that encompasses voice recordings from each speaker across many languages. This includes recordings in a single language, such as English or the speaker's mother tongue, as well as the identification of the speaker's proficiency in their second language. The presence of this multilingual database will streamline the research process, not only in the domain of speaker recognition, but also in broader applications such as multilingual language identification and the resilience of speaker models for forensic purposes.

e. Furthermore, the created database has considerable potential for use in identifying languages spoken in several languages, especially those spoken in India. Using the wide range of languages available in the database is expected to enhance the performance of the speaker identification system along with language identification, especially in relation to the extensive linguistic variety seen in India. In summary, the aim is to drive growth in speaker recognition and related domains, promoting innovation and advancement in the speech technology sector for different applications.

In conclusion, this study effort has been both very tough and quite gratifying. It has provided me with the chance to investigate and acquire a great lot of knowledge on speaker recognition. The hope is that it will motivate individuals passionate about speaker recognition, as well as other researchers, to contribute to the realization of the broader goal of integrating speaker identification technology into mainstream daily use.

Publications

The following articles are published as a direct outcome of the research conducted for this thesis.

In SCI Journals:

[1] Gaurav, Bhardwaj S, Agarwal R. Two-Tier Feature Extraction with Metaheuristics-Based Automated Forensic Speaker Verification Model. Electronics. 2023; 12(10):2342. https://doi.org/10.3390/electronics12102342 *(IF:2.9)*

[2] Gaurav, Bhardwaj S., & Agarwal R. An efficient speaker identification framework based on Mask R-CNN classifier parameter optimised using hosted cuckoo optimisation (HCO). J Ambient Intell Human Comput (2022). https://doi.org/10.1007/s12652-022-03828-7 *(IF:3.6)*

In Conferences:

[1] Gaurav, Bhardwaj S., & Agarwal R. "Refining Signal Quality for Elevating Speaker Diarization Performance," in Proceedings of the AdMet-2024 Conference, Gurugram University, Haryana, 7-8 March 2024

Unpublished work, under review:

[1] Gaurav, Bhardwaj S, Agarwal R. Automated Speaker Diarization System Using Arithmetic Optimization Algorithm with Deep Belief Network Model.

Summer school and workshops:

[1] Attended ISCA supported Summer School on Speech Signal Processing (S4P 2019) from 06th July to 10th July 2019 held at DA-IICT, Gandhinagar, Gujrat.

[2] Attended workshop for faculty on "Machine Learning and Deep Learning" from March 19th to March 24, 2018, conducted by IIT Guwahati at Gauhati University, Gauhati.

[3] Participated in the 6-week Thapar summer school on "Machine Learning and Deep Learning-2019" from June 03rd to July 12th, 2019 organized by the computer science and engineering department, TIET Patiala.

[4] Completed a course on "Neural Network and Machine learning" on April 15, 2020, on Coursera.

Bibliography

[1] S. Maes, H. Beigi, U. Chaudhari, and J. Sorensen, “IBM Model-Based and Frame-by-Frame Speaker-Recognition,” Jul. 1998.

[2] H. Beigi, S. Maes, U. Chaudhari, and J. Sorensen, *A hierarchical approach to large-scale speaker recognition*. 1999. doi: 10.21437/Eurospeech.1999-488.

[3] N. Singh, R. A. Khan, and R. Shree, “Applications of Speaker Recognition,” *Procedia Engineering*, vol. 38, pp. 3122–3126, 2012.

[4] D. A. Reynolds, “An overview of automatic speaker recognition technology,” in *2002 IEEE international conference on acoustics, speech, and signal processing*, IEEE, 2002, p. IV–4072.

[5] P. Kenny, T. Stafylakis, P. Ouellet, V. Gupta, and J. Alam, “Deep Neural Networks for extracting Baum-Welch statistics for Speaker Recognition,” in *The Speaker and Language Recognition Workshop (Odyssey 2014)*, ISCA, Jun. 2014, pp. 293–298. doi: 10.21437/Odyssey.2014-44.

[6] J. Wayman, A. Jain, D. Maltoni, and D. Maio, “An Introduction to Biometric Authentication Systems,” in *Biometric Systems: Technology, Design and Performance Evaluation*, J. Wayman, A. Jain, D. Maltoni, and D. Maio, Eds., London: Springer London, 2005, pp. 1–20. doi: 10.1007/1-84628-064-8_1.

[7] D. O’Shaughnessy, “Speaker recognition,” *IEEE ASSP Mag.*, vol. 3, no. 4, pp. 4–17, Oct. 1986, doi: 10.1109/MASSP.1986.1165388.

[8] P. Rose, “Forensic Speaker Identification”.

[9] D. Reynolds *et al.*, “The SuperSID project: exploiting high-level information for high-accuracy speaker recognition,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03).*, Hong Kong, China: IEEE, 2003, p. IV-784–7. doi: 10.1109/ICASSP.2003.1202760.

[10] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, “Fusing high- and low-level features for speaker recognition,” in *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, ISCA, Sep. 2003, pp. 2665–2668. doi: 10.21437/Eurospeech.2003-727.

[11] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Communication*, vol. 52, no. 1, pp. 12–40, Jan. 2010, doi: 10.1016/j.specom.2009.08.009.

[12] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing (2nd ed.)*. USA: Prentice-Hall, Inc., 1999.

[13] A. K. Singh and B. C. Pal, “Rate of Change of Frequency Estimation for Power Systems Using Interpolated DFT and Kalman Filter,” *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 2509–2517, Jul. 2019, doi: 10.1109/TPWRS.2018.2881151.

[14] J. Makhoul, “Linear prediction: A tutorial review,” *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975, doi: 10.1109/PROC.1975.9792.

[15] L. R. Rabiner and B.-H. Juang, “Fundamentals of speech recognition,” in *Prentice Hall signal processing series*, 1993. [Online]. Available: https://api.semanticscholar.org/CorpusID:7788300

[16] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990, doi: 10.1121/1.399423.

[17] T. Kinnunen, “Designing a speaker-discriminative adaptive filter bank for speaker recognition,” in *7th International Conference on Spoken Language Processing (ICSLP 2002)*, ISCA, Sep. 2002, pp. 2325–2328. doi: 10.21437/ICSLP.2002-630.

[18] K. K. Paliwal, “On the use of line spectral frequency parameters for speech recognition,” *Digital Signal Processing*, vol. 2, no. 2, pp. 80–87, Apr. 1992, doi: 10.1016/1051-2004(92)90028-W.

[19] C. Koniaris, S. Chatterjee, and W. B. Kleijn, “Selecting static and dynamic features using an advanced auditory model for speech recognition,” in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX, USA: IEEE, 2010, pp. 4342–4345. doi: 10.1109/ICASSP.2010.5495648.

[20] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, “Modeling prosodic dynamics for speaker recognition,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03).*, Hong Kong, China: IEEE, 2003, p. IV-788–91. doi: 10.1109/ICASSP.2003.1202761.
[21] S. Kajarekar *et al.*, “Speaker recognition using prosodic and lexical features,” in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, St Thomas, VI, USA: IEEE, 2003, pp. 19–24. doi: 10.1109/ASRU.2003.1318397.
[22] B. Peskin *et al.*, “Using prosodic and conversational features for high-performance speaker recognition: report from JHU WS’02,” in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03).*, Hong Kong, China: IEEE, 2003, p. IV-792–5. doi: 10.1109/ICASSP.2003.1202762.
[23] G. Doddington, “Speaker recognition based on idiolectal differences between speakers,” in *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, ISCA, Sep. 2001, pp. 2521–2524. doi: 10.21437/Eurospeech.2001-417.
[24] C. Müller, Ed., *Speaker Classification I: Fundamentals, Features, and Methods*, vol. 4343. in Lecture Notes in Computer Science, vol. 4343. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. doi: 10.1007/978-3-540-74200-5.
[25] T. Kitamura, K. Honda, and H. Takemoto, “Individual variation of the hypopharyngeal cavities and its acoustic effects,” *Acoust. Sci. & Tech.*, vol. 26, no. 1, pp. 16–26, 2005, doi: 10.1250/ast.26.16.
[26] E. Fisher, J. Tabrikian, and S. Dubnov, “Generalized Likelihood Ratio Test for Voiced-Unvoiced Decision in Noisy Speech Using the Harmonic Model,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 2, pp. 502–510, Mar. 2006, doi: 10.1109/TSA.2005.857806.
[27] J. Ramirez, J. C. Segura, C. Benitez, A. delaTorre, and A. J. Rubio, “A New Kullback–Leibler VAD for Speech Recognition in Noise,” *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 266–269, Feb. 2004, doi: 10.1109/LSP.2003.821762.
[28] J. Ramirez, J. C. Segura, C. Benitez, L. Garcia, and A. Rubio, “Statistical voice activity detection using a multiple observation likelihood ratio test,” *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, Oct. 2005, doi: 10.1109/LSP.2005.855551.
[29] L. R. Rabiner and M. R. Sambur, “An Algorithm for Determining the Endpoints of Isolated Utterances,” *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, Feb. 1975, doi: 10.1002/j.1538-7305.1975.tb02840.x.
[30] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999, doi: 10.1109/97.736233.
[31] O. Toledo-Ronen, “Speech Detection for Text-Dependent Speaker Verification”.
[32] F. Bimbot *et al.*, “A Tutorial on Text-Independent Speaker Verification,” *EURASIP J. Adv. Signal Process.*, vol. 2004, no. 4, p. 101962, Dec. 2004, doi: 10.1155/S1110865704310024.
[33] D. A. Reynolds, “Automatic speaker recognition using Gaussian mixture speaker models”.
[34] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995, doi: 10.1109/89.365379.
[35] P. Kumar and N. C. Sahu, “Export Potential of Climate Smart Goods in India: Evidence from the Poisson Pseudo Maximum Likelihood Estimator,” *The International Trade Journal*, vol. 35, no. 3, pp. 288–308, May 2021, doi: 10.1080/08853908.2021.1890652.
[36] N. Dehak, P. Dumouchel, and P. Kenny, “Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 7, pp. 2095–2103, Sep. 2007, doi: 10.1109/TASL.2007.902758.
[37] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-End Factor Analysis for Speaker Verification,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011, doi: 10.1109/TASL.2010.2064307.

[38] F. Richardson, D. Reynolds, and N. Dehak, “Deep Neural Network Approaches to Speaker and Language Recognition,” *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1671–1675, Oct. 2015, doi: 10.1109/LSP.2015.2420092.
[39] M. S. Nazir and C. Zhang, “Hybrid intelligent deep learning model for solar radiation forecasting using optimal variational mode decomposition and evolutionary deep belief network - Online sequential extreme learning machine,” *Journal of Building Engineering*, vol. 76, p. 107227, Oct. 2023, doi: 10.1016/j.jobe.2023.107227.
[40] D. A. Van Leeuwen and N. Brümmer, “An Introduction to Application-Independent Evaluation of Speaker Recognition Systems,” in *Speaker Classification I*, vol. 4343, C. Müller, Ed., in Lecture Notes in Computer Science, vol. 4343. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 330–353. doi: 10.1007/978-3-540-74200-5_19.
[41] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, “Speaker Diarization: A Review of Recent Research,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012, doi: 10.1109/TASL.2011.2125954.
[42] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, “A review of speaker diarization: Recent advances with deep learning,” *Computer Speech & Language*, vol. 72, p. 101317, Mar. 2022, doi: 10.1016/j.csl.2021.101317.
[43] E. Shriberg, A. Stolcke, and D. Baron, “Observations on overlap: findings and implications for automatic processing of multi-party conversation,” in *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, ISCA, Sep. 2001, pp. 1359–1362. doi: 10.21437/Eurospeech.2001-352.
[44] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, “Clustering speakers by their voices,” in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP ’98 (Cat. No.98CH36181)*, Seattle, WA, USA: IEEE, 1998, pp. 757–760. doi: 10.1109/ICASSP.1998.675375.
[45] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, Jan. 2000, doi: 10.1006/dspr.1999.0361.
[46] J. A. Bilmes, “A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models”.
[47] R. Mardhotillah, B. Dirgantoro, and C. Setianingsih, “Speaker Recognition for Digital Forensic Audio Analysis using Support Vector Machine,” in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia: IEEE, Dec. 2020, pp. 514–519. doi: 10.1109/ISRITI51436.2020.9315351.
[48] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, “Support vector machines for speaker and language recognition,” *Computer Speech & Language*, vol. 20, no. 2–3, pp. 210–229, Apr. 2006, doi: 10.1016/j.csl.2005.06.003.
[49] J. Zhang, “Realization and improvement algorithm of GMM - UBM model in voiceprint recognition,” in *2018 Chinese Control And Decision Conference (CCDC)*, Shenyang, China: IEEE, Jun. 2018, pp. 2989–2992. doi: 10.1109/CCDC.2018.8407636.
[50] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, “SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation,” in *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, Toulouse, France: IEEE, 2006, p. I-97-I–100. doi: 10.1109/ICASSP.2006.1659966.
[51] P. Kenny, “Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms,” 2006.
[52] O. Glembek, L. Burget, P. Matejka, M. Karafiat, and P. Kenny, “Simplification and optimization of i-vector extraction,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic: IEEE, May 2011, pp. 4516–4519. doi: 10.1109/ICASSP.2011.5947358.

[53] O. Ghahabi and J. Hernando, “Deep Learning Backend for Single and Multisession i-Vector Speaker Recognition,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 4, pp. 807–817, Apr. 2017, doi: 10.1109/TASLP.2017.2661705.
[54] S. J. D. Prince and J. H. Elder, “Probabilistic Linear Discriminant Analysis for Inferences About Identity,” in *2007 IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil: IEEE, 2007, pp. 1–8. doi: 10.1109/ICCV.2007.4409052.
[55] H. Dincer, “Pattern Recognition of Green Energy Innovation Investments Using a Modified Decision Support System,” *IEEE Access*, vol. 9, pp. 162006–162017, 2021, doi: 10.1109/ACCESS.2021.3133389.
[56] J. K. Basu, D. Bhattacharyya, and T. Kim, “Use of Artificial Neural Network in Pattern Recognition,” *International Journal of Software Engineering and Its Applications*, vol. 4, no. 2, 2010.
[57] R. McClanahan and P. L. De Leon, “Reducing computation in an i-vector speaker recognition system using a tree-structured universal background model,” *Speech Communication*, vol. 66, pp. 36–46, Feb. 2015, doi: 10.1016/j.specom.2014.07.003.
[58] M. McLaren, Y. Lei, and L. Ferrer, “Advances in deep neural network approaches to speaker recognition,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 4814–4818. doi: 10.1109/ICASSP.2015.7178885.
[59] I. Pollack, J. M. Pickett, and W. H. Sumby, “On the Identification of Speakers by Voice,” *The Journal of the Acoustical Society of America*, vol. 26, no. 3, pp. 403–406, Jun. 2005, doi: 10.1121/1.1907349.
[60] S. Pruzansky, “Pattern-Matching Procedure for Automatic Talker Recognition,” *The Journal of the Acoustical Society of America*, vol. 35, no. 3, pp. 354–358, Jul. 2005, doi: 10.1121/1.1918467.
[61] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, Jan. 2000, doi: 10.1006/dspr.1999.0361.
[62] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” *Speech Communication*, vol. 17, no. 1, pp. 91–108, Aug. 1995, doi: 10.1016/0167-6393(95)00009-D.
[63] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, “Speaker Diarization: A Review of Recent Research,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, Feb. 2012, doi: 10.1109/TASL.2011.2125954.
[64] G. Chaudhary, S. Srivastava, and S. Bhardwaj, “Feature Extraction Methods for Speaker Recognition: A Review,” *Int. J. Patt. Recogn. Artif. Intell.*, vol. 31, no. 12, p. 1750041, Dec. 2017, doi: 10.1142/S0218001417500410.
[65] B. S. Atal, “Automatic recognition of speakers from their voices,” *Proc. IEEE*, vol. 64, no. 4, pp. 460–475, 1976, doi: 10.1109/PROC.1976.10155.
[66] A. Khosravani and M. M. Homayounpour, “A PLDA approach for language and text independent speaker recognition,” *Computer Speech & Language*, vol. 45, pp. 457–474, Sep. 2017, doi: 10.1016/j.csl.2017.04.003.
[67] D. A. Reynolds, “An Overview of Automatic Speaker Recognition Technology·”.
[68] E. Bunge, “Automatic speaker recognition by computers,” in *ICASSP ’76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA: IEEE, 1976, pp. 738–738. doi: 10.1109/ICASSP.1976.1169970.
[69] P. Jesorsky, U. Hofker, and M. Talmi, “Extraction of speaker-specific features from spoken code sentences,” in *ICASSP ’78. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tulsa, Oklahoma: Institute of Electrical and Electronics Engineers, 1978, pp. 279–282. doi: 10.1109/ICASSP.1978.1170396.

[70] Alvin Martin and Mark Przybocki, “Speaker Recognition in a Multi-Speaker Environment,” European Conference on Speech Communication and Technology | 7th | | ISCA, Aalborg, DE, Sep. 2001. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=151526
[71] N. Singh, R. A. Khan, and R. Shree, “Applications of Speaker Recognition,” *Procedia Engineering*, vol. 38, pp. 3122–3126, 2012, doi: 10.1016/j.proeng.2012.06.363.
[72] C. D. Shaver and J. M. Acken, “A Brief Review of Speaker Recognition Technology”.
[73] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, “Robust Speaker Recognition in Noisy Conditions,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007, doi: 10.1109/TASL.2007.899278.
[74] B. S. Atal, “Text-Independent Speaker Recognition,” *The Journal of the Acoustical Society of America*, vol. 52, no. 1A_Supplement, pp. 181–181, Jul. 1972, doi: 10.1121/1.1982136.
[75] J. H. Han *et al.*, “Machine learning-based self-powered acoustic sensor for speaker recognition,” *Nano Energy*, vol. 53, pp. 658–665, Nov. 2018, doi: 10.1016/j.nanoen.2018.09.030.
[76] O. Rubanenko, “Optimal Value Determining Method of Parameters Compensating Devices of Bulk Electric Networks,” in *2022 IEEE 8th International Conference on Energy Smart Systems (ESS)*, Kyiv, Ukraine: IEEE, Oct. 2022, pp. 63–68. doi: 10.1109/ESS57819.2022.9969291.
[77] R. Venkatesan and A. Balaji Ganesh, “Binaural Classification-Based Speech Segregation and Robust Speaker Recognition System,” *Circuits Syst Signal Process*, vol. 37, no. 8, pp. 3383–3411, Aug. 2018, doi: 10.1007/s00034-017-0712-5.
[78] J. Sangeetha and T. Jayasankar, “A novel whispered speaker identification system based on extreme learning machine,” *Int J Speech Technol*, vol. 21, no. 1, pp. 157–165, Mar. 2018, doi: 10.1007/s10772-017-9488-z.
[79] L. Sun, T. Gu, K. Xie, and J. Chen, “Text-independent speaker identification based on deep Gaussian correlation supervector,” *Int J Speech Technol*, vol. 22, no. 2, pp. 449–457, Jun. 2019, doi: 10.1007/s10772-019-09618-5.
[80] S. Shon, H. Tang, and J. Glass, “Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model.” arXiv, Sep. 12, 2018. Accessed: Jan. 07, 2024. [Online]. Available: http://arxiv.org/abs/1809.04437
[81] N. N. An, N. Q. Thanh, and Y. Liu, “Deep CNNs With Self-Attention for Speaker Identification,” *IEEE Access*, vol. 7, pp. 85327–85337, 2019, doi: 10.1109/ACCESS.2019.2917470.
[82] A. B. Nassif, I. Shahin, S. Hamsa, N. Nemmour, and K. Hirose, “CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions,” *Applied Soft Computing*, vol. 103, p. 107141, May 2021, doi: 10.1016/j.asoc.2021.107141.
[83] M. Al-Qaderi, E. Lahamer, and A. Rad, “A Two-Level Speaker Identification System via Fusion of Heterogeneous Classifiers and Complementary Feature Cooperation,” *Sensors*, vol. 21, no. 15, p. 5097, Jul. 2021, doi: 10.3390/s21155097.
[84] S. Hourri, N. S. Nikolov, and J. Kharroubi, “Convolutional neural network vectors for speaker recognition,” *Int J Speech Technol*, vol. 24, no. 2, pp. 389–400, Jun. 2021, doi: 10.1007/s10772-021-09795-2.
[85] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 2, pp. 254–272, Apr. 1981, doi: 10.1109/TASSP.1981.1163530.
[86] N. S. Jayant, “Decision-Theoretic Approach to Speaker Verification,” *The Journal of the Acoustical Society of America*, vol. 51, no. 1A_Supplement, pp. 132–132, Aug. 2005, doi: 10.1121/1.1981418.
[87] A. E. Rosenberg, “Evaluation of an automatic speaker verification system over telephone lines,” *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S23–S23, Apr. 1975, doi: 10.1121/1.1995126.

[88] C. A. McGONEGAL and L. R. Rabiner, “Speaker Verification by Human Listeners over Several Speech Transmission Systems”.
[89] G. Baranov and I. O. Zaitsev, “S.M.A.R.T. Technologies for Transport Tests Networks, Exploitation and Repair Tools,” in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India: IEEE, Mar. 2021, pp. 621–625. doi: 10.1109/ICAIS50930.2021.9396055.
[90] P. Goli and M. R. Karami-mollaei, “Speech intelligibility improvement in noisy environments based on energy correlation in frequency bands,” *Digital Signal Processing*, vol. 62, pp. 238–248, Mar. 2017, doi: 10.1016/j.dsp.2016.12.005.
[91] K. Chen, “Towards better making a decision in speaker verification,” *Pattern Recognition*, vol. 36, no. 2, pp. 329–346, Feb. 2003, doi: 10.1016/S0031-3203(02)00034-1.
[92] M. Dua, A. Sadhu, A. Jindal, and R. Mehta, “A hybrid noise robust model for multireplay attack detection in Automatic speaker verification systems,” *Biomedical Signal Processing and Control*, vol. 74, p. 103517, Apr. 2022, doi: 10.1016/j.bspc.2022.103517.
[93] A. Larcher, K. A. Lee, B. Ma, and H. Li, “Text-dependent speaker verification: Classifiers, databases and RSR2015,” *Speech Communication*, vol. 60, pp. 56–77, May 2014, doi: 10.1016/j.specom.2014.03.001.
[94] J. Wayman, N. Orlans, Q. Hu, F. Goodman, A. Ulrich, and V. Valencia, “Face, Iris, Ear, Voice, and Handwriter Recognition March 2009; v1.3”.
[95] A. K. H. Al-Ali, D. Dean, B. Senadji, V. Chandran, and G. R. Naik, “Enhanced Forensic Speaker Verification Using a Combination of DWT and MFCC Feature Warping in the Presence of Noise and Reverberation Conditions,” *IEEE Access*, vol. 5, pp. 15400–15413, 2017, doi: 10.1109/ACCESS.2017.2728801.
[96] S. Huang, H. Dang, R. Jiang, Y. Hao, C. Xue, and W. Gu, “Multi-Layer Hybrid Fuzzy Classification Based on SVM and Improved PSO for Speech Emotion Recognition,” *Electronics*, vol. 10, no. 23, p. 2891, Nov. 2021, doi: 10.3390/electronics10232891.
[97] A. Kalam, Md. Alamgir Hossain, R. Manojkumar, and N. Saeed, “Optimizing PV and Battery Energy Storage Systems for Peak Demand Reduction and Cost Savings in Malaysian Commercial Buildings,” in *2023 IEEE IAS Global Conference on Emerging Technologies (GlobConET)*, London, United Kingdom: IEEE, May 2023, pp. 1–6. doi: 10.1109/GlobConET56651.2023.10149980.
[98] M. Swain, B. Maji, P. Kabisatpathy, and A. Routray, “A DCRNN-based ensemble classifier for speech emotion recognition in Odia language,” *Complex Intell. Syst.*, vol. 8, no. 5, pp. 4237–4249, Oct. 2022, doi: 10.1007/s40747-022-00713-w.
[99] S. Saleem, F. Subhan, N. Naseer, A. Bais, and A. Imtiaz, “Forensic speaker recognition: A new method based on extracting accent and language information from short utterances,” *Forensic Science International: Digital Investigation*, vol. 34, p. 300982, Sep. 2020, doi: 10.1016/j.fsidi.2020.300982.
[100] K. J. Devi, N. H. Singh, and K. Thongam, “Automatic Speaker Recognition from Speech Signals Using Self Organizing Feature Map and Hybrid Neural Network,” *Microprocessors and Microsystems*, vol. 79, p. 103264, Nov. 2020, doi: 10.1016/j.micpro.2020.103264.
[101] F. Teixeira, A. Abad, B. Raj, and I. Trancoso, “Towards End-to-End Private Automatic Speaker Recognition,” in *Interspeech 2022*, Sep. 2022, pp. 2798–2802. doi: 10.21437/Interspeech.2022-10672.
[102] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, “A review of speaker diarization: Recent advances with deep learning,” *Computer Speech & Language*, vol. 72, p. 101317, Mar. 2022, doi: 10.1016/j.csl.2021.101317.
[103] V. Khoma, Y. Khoma, V. Brydinskyi, and A. Konovalov, “Development of Supervised Speaker Diarization System Based on the PyAnnote Audio Processing Library,” *Sensors*, vol. 23, no. 4, p. 2082, Feb. 2023, doi: 10.3390/s23042082.

[104] M. Kumar, S. H. Kim, C. Lord, and S. Narayanan, “Improving speaker diarization for naturalistic child-adult conversational interactions using contextual information,” *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. EL196–EL200, Feb. 2020, doi: 10.1121/10.0000736.
[105] Y. Prokopalo, M. Shamsi, L. Barrault, S. Meignier, and A. Larcher, “Active Correction for Incremental Speaker Diarization of a Collection with Human in the Loop,” *Applied Sciences*, vol. 12, no. 4, p. 1782, Feb. 2022, doi: 10.3390/app12041782.
[106] Z. Huang, M. Delcroix, L. P. Garcia, S. Watanabe, D. Raj, and S. Khudanpur, “Joint speaker diarization and speech recognition based on region proposal networks,” *Computer Speech & Language*, vol. 72, p. 101316, Mar. 2022, doi: 10.1016/j.csl.2021.101316.
[107] R. Ahmad, S. Zubair, and H. Alquhayz, “Speech Enhancement for Multimodal Speaker Diarization System,” *IEEE Access*, vol. 8, pp. 126671–126680, 2020, doi: 10.1109/ACCESS.2020.3007312.
[108] Q. Lin, Y. Hou, and M. Li, “Self-Attentive Similarity Measurement Strategies in Speaker Diarization,” in *Interspeech 2020*, ISCA, Oct. 2020, pp. 284–288. doi: 10.21437/Interspeech.2020-1908.
[109] V.-T. Tran and W.-H. Tsai, “Speaker Identification in Multi-Talker Overlapping Speech Using Neural Networks,” *IEEE Access*, vol. 8, pp. 134868–134879, 2020, doi: 10.1109/ACCESS.2020.3009987.
[110] K. A. Kamiński and A. P. Dobrowolski, “Automatic Speaker Recognition System Based on Gaussian Mixture Models, Cepstral Analysis, and Genetic Selection of Distinctive Features,” *Sensors*, vol. 22, no. 23, p. 9370, Dec. 2022, doi: 10.3390/s22239370.
[111] J. M. Naik, “Speaker verification: a tutorial,” *IEEE Commun. Mag.*, vol. 28, no. 1, pp. 42–48, Jan. 1990, doi: 10.1109/35.46670.
[112] G. R. Doddington, “Speaker recognition—Identifying people by their voices,” *Proc. IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985, doi: 10.1109/PROC.1985.13345.
[113] S. Furui and M. M. Sondhi, “Advances in Speech Signal Processing,” 1991. [Online]. Available: https://api.semanticscholar.org/CorpusID:58146772
[114] “Introduction to the Special Section on Sound Scene and Event Analysis,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 6, pp. 1169–1171, Jun. 2017, doi: 10.1109/TASLP.2017.2699334.
[115] C. S. Greenberg, L. P. Mason, S. O. Sadjadi, and D. A. Reynolds, “Two decades of speaker recognition evaluation at the national institute of standards and technology,” *Computer Speech & Language*, vol. 60, p. 101032, Mar. 2020, doi: 10.1016/j.csl.2019.101032.
[116] R. Jahangir *et al.*, “Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network,” *IEEE Access*, vol. 8, pp. 32187–32202, 2020, doi: 10.1109/ACCESS.2020.2973541.
[117] I. Bisio, C. Garibotto, A. Grattarola, F. Lavagetto, and A. Sciarrone, “Smart and Robust Speaker Recognition for Context-Aware In-Vehicle Applications,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8808–8821, Sep. 2018, doi: 10.1109/TVT.2018.2849577.
[118] M. El Ayadi, A.-K. S.O. Hassan, A. Abdel-Naby, and O. A. Elgendy, “Text-independent speaker identification using robust statistics estimation,” *Speech Communication*, vol. 92, pp. 52–63, Sep. 2017, doi: 10.1016/j.specom.2017.05.005.
[119] S. Hourri and J. Kharroubi, “A deep learning approach for speaker recognition,” *Int J Speech Technol*, vol. 23, no. 1, pp. 123–131, Mar. 2020, doi: 10.1007/s10772-019-09665-y.
[120] K. J. Devi and K. Thongam, “Automatic speaker recognition with enhanced swallow swarm optimization and ensemble classification model from speech signals,” *J Ambient Intell Human Comput*, Jul. 2019, doi: 10.1007/s12652-019-01414-y.
[121] M. Ravanelli and Y. Bengio, “Speaker Recognition from Raw Waveform with SincNet.” arXiv, Aug. 09, 2019. Accessed: Feb. 02, 2024. [Online]. Available: http://arxiv.org/abs/1808.00158

[122] F. H. Shajin and P. Rajesh, “Trusted Secure Geographic Routing Protocol: outsider attack detection in mobile ad hoc networks by adopting trusted secure geographic routing protocol,” *International Journal of Pervasive Computing and Communications*, vol. 18, no. 5, pp. 603–621, Jan. 2022, doi: 10.1108/IJPCC-09-2020-0136.
[123] S. Nainan and V. Kulkarni, “Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN,” *Int J Speech Technol*, vol. 24, no. 4, pp. 809–822, Dec. 2021, doi: 10.1007/s10772-020-09771-2.
[124] J. Villalba *et al.*, “State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations,” *Computer Speech & Language*, vol. 60, p. 101026, Mar. 2020, doi: 10.1016/j.csl.2019.101026.
[125] P. Rajesh and F. H. Shajin, “A Multi-Objective Hybrid Algorithm for Planning Electrical Distribution System,” *EJEE*, vol. 22, no. 4–5, pp. 224–509, Oct. 2020, doi: 10.18280/ejee.224-509.
[126] S. A. El-Moneim, M. A. Nassar, M. I. Dessouky, N. A. Ismail, A. S. El-Fishawy, and F. E. Abd El-Samie, “Text-independent speaker recognition using LSTM-RNN and speech enhancement,” *Multimed Tools Appl*, vol. 79, no. 33–34, pp. 24013–24028, Sep. 2020, doi: 10.1007/s11042-019-08293-7.
[127] S. M. Jagdale, A. A. Shinde, and J. S. Chitode, “Robust Speaker Recognition Based on Low-Level- and Prosodic-Level-Features,” in *Advances in Data Sciences, Security and Applications*, V. Jain, G. Chaudhary, M. C. Taplamacioglu, and M. S. Agarwal, Eds., Singapore: Springer Singapore, 2020, pp. 267–274.
[128] V. Reddy and G. Prakash, “Enhanced key establishment technique for secure data access in cloud,” in *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, GHAZIABAD, India: IEEE, Sep. 2019, pp. 1–4. doi: 10.1109/ICICT46931.2019.8977720.
[129] Mahesh Kumar Thota, Francis H Shajin, and P. Rajesh, “Survey on software defect prediction techniques,” *International Journal of Applied Science and Engineering*, vol. 17, no. 4, Dec. 2020, doi: 10.6703/IJASE.202012_17(4).331.
[130] M. Jessen, J. Bortlík, P. Schwarz, and Y. A. Solewicz, “Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01),” *Speech Communication*, vol. 111, pp. 22–28, Aug. 2019, doi: 10.1016/j.specom.2019.05.002.
[131] M. Geravanchizadeh, E. Forouhandeh, and M. Bashirpour, “Feature compensation based on the normalization of vocal tract length for the improvement of emotion-affected speech recognition,” *J AUDIO SPEECH MUSIC PROC.*, vol. 2021, no. 1, p. 31, Dec. 2021, doi: 10.1186/s13636-021-00216-5.
[132] M. C. Madhavi and H. A. Patil, “Vocal Tract Length Normalization using a Gaussian mixture model framework for query-by-example spoken term detection,” *Computer Speech & Language*, vol. 58, pp. 175–202, Nov. 2019, doi: 10.1016/j.csl.2019.03.005.
[133] U. Kumaran, S. Radha Rammohan, S. M. Nagarajan, and A. Prathik, “Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN,” *Int J Speech Technol*, vol. 24, no. 2, pp. 303–314, Jun. 2021, doi: 10.1007/s10772-020-09792-x.
[134] S. S. Therese and C. Lingam, “A linear visual assessment tendency based clustering with power normalized cepstral coefficients for audio signal recognition system,” *J Ambient Intell Human Comput*, Dec. 2017, doi: 10.1007/s12652-017-0653-7.
[135] C. Nicolini, G. Forcellini, L. Minati, and A. Bifone, “Scale-resolved analysis of brain functional connectivity networks with spectral entropy,” *NeuroImage*, vol. 211, p. 116603, May 2020, doi: 10.1016/j.neuroimage.2020.116603.

[136] M. A. Mellal, A. Frik, and R. Boutiche, “Chapter 1 - Reliability optimization of power plant safety system using grey wolf optimizer and shuffled frog-leaping algorithm,” in *Nature-Inspired Computing Paradigms in Systems*, M. A. Mellal and M. G. Pecht, Eds., Academic Press, 2021, pp. 1–13. doi: 10.1016/B978-0-12-823749-6.00008-8.
[137] R. Sharma and P. Gaur, “Performance evaluation of cuckoo search algorithm based FOPID controllers applied to a robotic manipulator with actuator,” in *2015 International Conference on Advances in Computer Engineering and Applications*, Ghaziabad, India: IEEE, Mar. 2015, pp. 356–363. doi: 10.1109/ICACEA.2015.7164730.
[138] B. Zagagy, M. Herman, and O. Levi, “ACKEM: Automatic Classification, Using KNN Based Ensemble Modeling,” in *Advances in Information and Communication*, K. Arai, Ed., Cham: Springer International Publishing, 2021, pp. 536–557.
[139] C. Chen, W. Wang, Y. He, and J. Han, “A bilevel framework for joint optimization of session compensation and classification for speaker identification,” *Digital Signal Processing*, vol. 89, pp. 104–115, Jun. 2019, doi: 10.1016/j.dsp.2019.03.008.
[140] I. Shahin, A. B. Nassif, and S. Hamsa, “Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments,” *Neural Comput & Applic*, vol. 32, no. 7, pp. 2575–2587, Apr. 2020, doi: 10.1007/s00521-018-3760-2.
[141] B. Chaudhuri and B. C. Pal, “Robust Damping of Multiple Swing Modes Employing Global Stabilizing Signals With a TCSC,” *IEEE Trans. Power Syst.*, vol. 19, no. 1, pp. 499–506, Feb. 2004, doi: 10.1109/TPWRS.2003.821463.
[142] D. A. Reynolds, “Automatic Speaker Recognition: Current Approaches and Future Trends •1,” 2001. [Online]. Available: https://api.semanticscholar.org/CorpusID:14594203
[143] Machado, Vieira Filho, and De Oliveira, “Forensic Speaker Verification Using Ordinary Least Squares,” *Sensors*, vol. 19, no. 20, p. 4385, Oct. 2019, doi: 10.3390/s19204385.
[144] Z. Wang, W. Xia, and J. H. L. Hansen, “Cross-domain Adaptation with Discrepancy Minimization for Text-independent Forensic Speaker Verification.” arXiv, Sep. 09, 2020. Accessed: Feb. 04, 2024. [Online]. Available: http://arxiv.org/abs/2009.02444
[145] I. Stefanus, R. S. J. Sarwono, and M. I. Mandasari, “GMM based automatic speaker verification system development for forensics in Bahasa Indonesia,” in *2017 5th International Conference on Instrumentation, Control, and Automation (ICA)*, Yogyakarta, Indonesia: IEEE, Aug. 2017, pp. 56–61. doi: 10.1109/ICA.2017.8068413.
[146] M. Algabri, H. Mathkour, M. A. Bencherif, M. Alsulaiman, and M. A. Mekhtiche, “Automatic Speaker Recognition for Mobile Forensic Applications,” *Mobile Information Systems*, vol. 2017, pp. 1–6, 2017, doi: 10.1155/2017/6986391.
[147] “Forensic Linguistic Inquiry into the Validity of F0 as Discriminatory Potential in the System of Forensic Speaker Verification,” *JFSCI*, vol. 5, no. 3, Sep. 2017, doi: 10.19080/JFSCI.2017.05.555664.
[148] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” *Computer Speech & Language*, vol. 60, p. 101027, Mar. 2020, doi: 10.1016/j.csl.2019.101027.
[149] M. S. Athulya and P. S. Sathidevi, “Speaker verification from codec distorted speech for forensic investigation through serial combination of classifiers,” *Digital Investigation*, vol. 25, pp. 70–77, Jun. 2018, doi: 10.1016/j.diin.2018.03.005.
[150] R. González Hautamäki, M. Sahidullah, V. Hautamäki, and T. Kinnunen, “Acoustical and perceptual study of voice disguise by age modification in speaker verification,” *Speech Communication*, vol. 95, pp. 1–15, Dec. 2017, doi: 10.1016/j.specom.2017.10.002.
[151] A. Poddar, M. Sahidullah, and G. Saha, “Speaker verification with short utterances: a review of challenges, trends and opportunities,” *IET biom.*, vol. 7, no. 2, pp. 91–101, Mar. 2018, doi: 10.1049/iet-bmt.2017.0065.

[152] S. Susanto and D. S. Nanda, “Analysing Forensic Speaker Verification by Utilizing Artificial Neural Network:,” presented at the International Congress of Indonesian Linguistics Society (KIMLI 2021), Makassar, Indonesia, 2021. doi: 10.2991/assehr.k.211226.026.
[153] S. Saif, P. Das, S. Biswas, M. Khari, and V. Shanmuganathan, “HIIDS: Hybrid intelligent intrusion detection system empowered with machine learning and metaheuristic algorithms for application in IoT based healthcare,” *Microprocessors and Microsystems*, p. 104622, Aug. 2022, doi: 10.1016/j.micpro.2022.104622.
[154] H. Gao, M. Hu, T. Gao, and R. Cheng, “Robust detection of median filtering based on combined features of difference image,” *Signal Processing: Image Communication*, vol. 72, pp. 126–133, Mar. 2019, doi: 10.1016/j.image.2018.12.014.
[155] Z. Ma and E. Fokoué, “Accent Recognition for Noisy Audio Signals,” *SJC*, vol. 8, no. 2, pp. 169–182, Apr. 2015, doi: 10.55630/sjc.2014.8.169-182.
[156] N. Dong, L. Zhao, C. H. Wu, and J. F. Chang, “Inception v3 based cervical cell classification combined with artificially extracted features,” *Applied Soft Computing*, vol. 93, p. 106311, Aug. 2020, doi: 10.1016/j.asoc.2020.106311.
[157] H. Dong, Y. Xu, X. Li, Z. Yang, and C. Zou, “An improved antlion optimizer with dynamic random walk and dynamic opposite learning,” *Knowledge-Based Systems*, vol. 216, p. 106752, Mar. 2021, doi: 10.1016/j.knosys.2021.106752.
[158] A. Singh, “Selection of Hidden Layer Neurons and Best Training Method for FFNN in Application of Long Term Load Forecasting,” *Journal of Electrical Engineering*, vol. 63, no. 3, pp. 153–161, May 2012, doi: 10.2478/v10187-012-0023-9.
[159] Y. Zhang, R. Xiong, H. He, and M. G. Pecht, “Long Short-Term Memory Recurrent Neural Network for Remaining Useful Life Prediction of Lithium-Ion Batteries,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 5695–5705, Jul. 2018, doi: 10.1109/TVT.2018.2805189.
[160] X. Anguera, C. Wooters, and J. Hernando, “Acoustic Beamforming for Speaker Diarization of Meetings,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007, doi: 10.1109/TASL.2007.902460.
[161] P. Kenny, D. Reynolds, and F. Castaldo, “Diarization of Telephone Conversations Using Factor Analysis,” *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 6, pp. 1059–1070, Dec. 2010, doi: 10.1109/JSTSP.2010.2081790.
[162] J. F. Bonastre, P. M. Bousquet, D. Matrouf, and X. Anguera, “Discriminant binary data representation for speaker recognition,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic: IEEE, May 2011, pp. 5284–5287. doi: 10.1109/ICASSP.2011.5947550.
[163] M. Diez, L. Burget, F. Landini, and J. Cernocky, “Analysis of Speaker Diarization Based on Bayesian HMM With Eigenvoice Priors,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 355–368, 2020, doi: 10.1109/TASLP.2019.2955293.
[164] Q. Lin, Y. Hou, and M. Li, “Self-Attentive Similarity Measurement Strategies in Speaker Diarization,” in *Interspeech 2020*, ISCA, Oct. 2020, pp. 284–288. doi: 10.21437/Interspeech.2020-1908.
[165] H. Bredin *et al.*, “pyannote.audio: neural building blocks for speaker diarization.” arXiv, Nov. 04, 2019. Accessed: Feb. 05, 2024. [Online]. Available: http://arxiv.org/abs/1911.01255
[166] D. Raj, Z. Huang, and S. Khudanpur, “Multi-class Spectral Clustering with Overlaps for Speaker Diarization.” arXiv, Nov. 05, 2020. Accessed: Feb. 04, 2024. [Online]. Available: http://arxiv.org/abs/2011.02900
[167] L. Fürer, N. Schenk, V. Roth, M. Steppan, K. Schmeck, and R. Zimmermann, “Supervised Speaker Diarization Using Random Forests: A Tool for Psychotherapy Process Research,” *Front. Psychol.*, vol. 11, p. 1726, Jul. 2020, doi: 10.3389/fpsyg.2020.01726.
[168] J. M. Coria, H. Bredin, S. Ghannay, and S. Rosset, “Continual Self-Supervised Domain Adaptation for End-to-End Speaker Diarization,” in *2022 IEEE Spoken Language Technology*

Workshop (SLT), Doha, Qatar: IEEE, Jan. 2023, pp. 626–632. doi: 10.1109/SLT54892.2023.10023195.
[169] Y. Takashima, Y. Fujita, S. Watanabe, S. Horiguchi, P. Garcia, and K. Nagamatsu, "End-to-End Speaker Diarization Conditioned on Speech Activity and Overlap Detection," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China: IEEE, Jan. 2021, pp. 849–856. doi: 10.1109/SLT48900.2021.9383555.
[170] Q. Chen and A. Brandt, "Speakership, recipiency and the interactional space: Cases of 'Next-speaker self-selects' in multiparty university student meetings," *Journal of Pragmatics*, vol. 180, pp. 54–71, Jul. 2021, doi: 10.1016/j.pragma.2021.04.012.
[171] X. Zhang, W. Wang, and P. Zhang, "Speaker Diarization System Based on DPCA Algorithm for Fearless Steps Challenge Phase-2," in *Interspeech 2020*, ISCA, Oct. 2020, pp. 2602–2606. doi: 10.21437/Interspeech.2020-1666.
[172] Q. Li, F. L. Kreyssig, C. Zhang, and P. C. Woodland, "Discriminative Neural Clustering for Speaker Diarisation." arXiv, Nov. 23, 2020. Accessed: Feb. 04, 2024. [Online]. Available: http://arxiv.org/abs/1910.09703
[173] J.-T. Chien and S. Luo, "Self-Supervised Learning for Online Speaker Diarization," 2021.
[174] P. Singh and J. S. Lather, "Artificial neural network-based dynamic power management of a DC microgrid: a hardware-in-loop real-time verification," *International Journal of Ambient Energy*, vol. 43, no. 1, pp. 1730–1738, Dec. 2022, doi: 10.1080/01430750.2020.1720811.
[175] T. Bosch, "Machine Learning for Automatic Signer Diarization".
[176] M. Shamsi *et al.*, "Towards lifelong human assisted speaker diarization," *Computer Speech & Language*, vol. 77, p. 101437, Jan. 2023, doi: 10.1016/j.csl.2022.101437.
[177] B. Mouaz, B. H. Abderrahim, and E. Abdelmajid, "Speech Recognition of Moroccan Dialect Using Hidden Markov Models," *Procedia Computer Science*, vol. 151, pp. 985–991, 2019, doi: 10.1016/j.procs.2019.04.138.
[178] C. Li, Z. Ding, J. Yi, Y. Lv, and G. Zhang, "Deep Belief Network Based Hybrid Model for Building Energy Consumption Prediction," *Energies*, vol. 11, no. 1, p. 242, Jan. 2018, doi: 10.3390/en11010242.
[179] L. Abualigah, A. Diabat, S. Mirjalili, M. Abd Elaziz, and A. H. Gandomi, "The Arithmetic Optimization Algorithm," *Computer Methods in Applied Mechanics and Engineering*, vol. 376, p. 113609, Apr. 2021, doi: 10.1016/j.cma.2020.113609.
[180] K. G. Dhal, B. Sasmal, A. Das, S. Ray, and R. Rai, "A Comprehensive Survey on Arithmetic Optimization Algorithm," *Arch Computat Methods Eng*, vol. 30, no. 5, pp. 3379–3404, Jun. 2023, doi: 10.1007/s11831-023-09902-3.
[181] P. R. Prakash, D. Anuradha, J. Iqbal, M. G. Galety, R. Singh, and S. Neelakandan, "A novel convolutional neural network with gated recurrent unit for automated speech emotion recognition and classification," *Journal of Control and Decision*, vol. 10, no. 1, pp. 54–63, Jan. 2023, doi: 10.1080/23307706.2022.2085198.
[182] C. Al-Atroshi, J. Rene Beulah, K. K. Singamaneni, C. Pretty Diana Cyril, S. Neelakandan, and S. Velmurugan, "Automated speech based evaluation of mild cognitive impairment and Alzheimer's disease detection using with deep belief network model," *International Journal of Healthcare Management*, pp. 1–11, doi: 10.1080/20479700.2022.2097764.
[183] N. Ryant *et al.*, "The Second DIHARD Diarization Challenge: Dataset, task, and baselines." arXiv, Jun. 18, 2019. Accessed: Feb. 05, 2024. [Online]. Available: http://arxiv.org/abs/1906.07839
[184] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch, and S. Khudanpur, "Improved Speech-to-Text Translation with the Fisher and Callhome Spanish–English Speech Translation Corpus," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, Dec. 2013.
[185] "A hybrid HXPLS-TMFCC parameterization and DCNN-SFO clustering-based speaker diarization system.pdf."

[186] P. Singh and S. Ganapathy, “Deep Self-Supervised Hierarchical Clustering for Speaker Diarization,” in *Interspeech 2020*, Oct. 2020, pp. 294–298. doi: 10.21437/Interspeech.2020-2297.
[187] M. Pal *et al.*, “Meta-learning with Latent Space Clustering in Generative Adversarial Network for Speaker Diarization.” arXiv, Jul. 19, 2020. Accessed: Feb. 04, 2024. [Online]. Available: http://arxiv.org/abs/2007.09635
[188] V. Subba Ramaiah and R. Rajeswara Rao, “Speaker diarization system using HXLPS and deep neural network,” *Alexandria Engineering Journal*, vol. 57, no. 1, pp. 255–266, Mar. 2018, doi: 10.1016/j.aej.2016.12.009.
[189] V. Subba Ramaiah and R. Rajeswara Rao, “A novel approach for speaker diarization system using TMFCC parameterization and Lion optimization,” *J. Cent. South Univ.*, vol. 24, no. 11, pp. 2649–2663, Nov. 2017, doi: 10.1007/s11771-017-3678-3.

Glossary

- Averager: A difference equation to approximate a low-pass filter.
- (Arithmetical) mean: The proper statistical term for average.
- Cepstrum: A very common parameter used in automatic speaker recognition, one effect of which is to smooth the spectrum.
- Cepstrum domain: A sample domain which is obtained when applying frequency analysis to the frequency domain.
- Differentiator: A difference equation to approximate a high-pass filter
- Divergence: A measure of the distance between two statistical models.
- DNA Features: Speaker specific features extracted using the deep network architecture.
- EER: Equal Error Rate, a special point on the ROC curve, where the miss ratio is equal to the false alarm ratio.
- False negative: In speaker recognition, deciding that two speech samples have come from different speakers when, in fact, they are from the same speaker.
- False positive: In speaker recognition, deciding that two speech samples have come from the same speaker when, in fact, they are from different speakers.
- Fast Fourier transform: a common method of spectral analysis in acoustic phonetics.
- Filter Bank: A set of filters to cover the whole range of frequency found in the signal.
- Formants: The higher frequency boundaries for a voiced part of speech.
- Formant bandwidth: an acoustic parameter that reflects the degree to which acoustic energy is absorbed in the vocal tract during speech.
- Fourier Transformation: A method to decompose a real life signal into a number of sinusoids.
- Frequency: The number of times the signal repeats itself in a time unit.
- Fundamental Frequency: The frequency at which the vocal cords vibrate.
- Gaussian Mixture Model: A mixture of more than one normal distribution, combined with different weights.
- Hertz (or Hz): Unit for quantifying frequency: so many times per second: 100 Hz, for example, means a frequency of one hundred times per second.
- Liner Predictive Coding: Compression technique using linear combinations and residual.
- Mel-Scale: A non-linear scale based on human frequency perception.
- Nyquist Frequency: The highest sampling frequency that can faithfully preserve the original signal.
- Period: The time a periodic signal needs to repeat itself.
- Periodic Signal: A signal that repeat itself every period T.
- Phoneme: a unit of linguistic analysis: the name for a contrastive sound in a language.
- Phonemics: the study of how speech sounds function contrastively, to distinguish words in a given language.
- Pitch: The human perception of the fundamental frequency.
- ROC Curves: Receiver Operating Characteristic, a plot to show the effect of different parameter values in a parameterised method.
- Signal Filtering: Affecting certain frequency domains of the signal.
- Signal Windowing: Extracting parts of the signal in the time domain.

- Source/filter model: A simple method to model the human voice production by using a source (base frequency generator) and a filter (filter bank to shape the final voice).
- Spectrum: the result of an acoustic analysis showing how much energy is present at what frequencies in a given amount of speech.
- Speaker Clustering: Group different speech segments according the speaker.
- Speaker Detection: Correctly flag the speeches, out of a set of speech, of a certain speaker.
- Speaker Diarization: Detect who is speaking when on a stream.
- Speaker Identification: Correctly identify the speaker, out of a certain set, of a certain speech.
- Speaker Modelling: Different techniques used to build a model to retain enough information about the speaker identity.
- Speaker Recognition SR: Any task related to automatically deal with the speaker identity.
- Speaker Segmentation: Automatically identify turn points of speakers in a large stream.
- Speaker Specific Features: A deep network architecture learned features for maximum discrimination among speakers.
- Speaker Verification: Correctly verify that two segments are uttered by the same speaker.
- Voiceprint: another name for spectrogram. Usually avoided because of its association with voiceprint identification.