[1]Neelam Nehra
[2]Pardeep Sangwan
[3]Neelu Trivedi

A Novel, Robust and Deep Learning based Speaker Recognition System using Three Different Datasets

***Abstract*: -** Speaker recognition methods have become increasingly popular across various domains like security, domestic services, smart terminals, speech communications and access control. However, current applications face challenges in accurately recognising speakers from short speech segments, common in modern interactive devices like smartphones and smart speakers. This paper introduces a novel, highly accurate and robust approach to address this issue by leveraging Convolutional Neural Networks (CNN) and LSTM (Long-short term memory- Recurrent Neural Networks (RNN). Three different databases, namely the SITW 2016, NIST 2008, and TIMIT, are used to evaluate the system performance of the data during different training and testing durations. According to the experimental results, our model LSTM-RNN with temporal learning and memory features performs significantly better than CNN, particularly when compared to short utterance durations. The proposed model presents the classification accuracy of 84.3%, 95.09%, 94% for 10s and, 85.4%, 96.47%, 95.24% for 20s training duration and for TIMIT, SITW 2016, and NIST 2008 datasets, respectively.

Keywords: Convolutional Neural Networks (CNN), Long short-term memory (LSTM) approach, Recurrent Neural Networks (RNN), TIMIT database, Additive Gaussian noise

I. INTRODUCTION

Biometric identification, or biometrics, is identifying people using their unique characteristics. Compared to more conventional techniques like using passwords, passports, and keys, all at risk of theft or forgery, it provides a higher level of security. Biometric technologies use behavioural features such as handwriting, gait and keystrokes as well as physiological characteristics like facial, fingerprint, and iris features [1]. A biometric system gathers information from an individual, extracts relevant features, and then compares those features with models that are stored within a database.

In human interaction, voice is sometimes taken for granted, yet it is increasingly important as a biometric trait. Speaker identification is essential for various services, such as voice dialling, banking, and security access. Speech signals can be used to verify a user's identification using automatic speaker recognition, providing a wide range of services [2].

Two fundamental applications of speaker recognition systems are Speaker identification and speaker verification. Speaker verification confirms the claimed identity, while speaker identification uses voice recognition from a collection of speaker models to identify specific individuals. These tasks can be either text-independent, allowing the use of any spoken phrase, or text-dependent, requiring specific phrases [3].

The speaker's health should have less influence on feature extraction in speaker identification. This research aims to increase the identification accuracy by combining the ELM with an i-vector to develop a robust speaker identification system. Additionally, research is carried out on the impact of non-stationary noise (NSN), and performance is evaluated using different fusion techniques. A comprehensive overview based on biometric fusion to examine various fusion methods. Moreover, recent research contributions are employed in the fields of speaker identification and fusion techniques [4-8].

To increase speaker identification accuracy in this instance, the literature suggests score fusion between the inverse MFCC features and the Mel-Frequency Cepstral Coefficients (MFCC)[9]. GMM was used to examine 120 speakers from dialect regions one and four of the TIMIT database; however, evaluation in noisy environments was unavailable [10].

[1]*Neelam Nehra: IFTM University Moradabad, MSIT New Delhi

[2]Pardeep Sangwan, MSIT New Delhi

[3]Neelu Trivedi: IFTM University Moradabad

A robust speaker identification method was proposed based on responses from an auditory periphery model. The effectiveness of this method for speaker identification was compared to the identification outcomes for the features of frequency domain linear prediction (FDLP), Gamma-tone Filter Cepstral Coefficient (GFCC), and MFCC. Three of the examined datasets—TIMIT, TIDIGT, and YOHO—were employed for text-independent study [11].

Due to the increase in the number of interactive gadgets like smart speakers and smartphones, speaker recognition, particularly with short utterances, has increased. Large amounts of speech data are needed for training in traditional speaker recognition techniques, which can be challenging in real-world environments. Additionally, present systems frequently fail to deal with short utterances, which is common in adverse surroundings [12-14].

Advanced speaker recognition systems use deep learning methods like Deep Neural Networks (DNNs) to overcome these challenges. In many types of pattern recognition applications, such as speaker recognition, DNN models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) [15-17] have shown promising performance.

This work proposes that DNN architecture employ short utterances for speaker identification. We use both CNN and RNN models for speaker identification tasks and provide a Deep Neural Network architecture based on these models. We apply Cepstral Mean and Variance Normalized coefficients (CMVNC) to improve system performance further.

Three databases are used to evaluate the proposed system: the widely used and well-known Training Set Part 2 of the 2008 NIST Speaker Recognition Evaluation database (NIST 2008) [18], Texas Instruments and Massachusetts Institute of Technology (TIMIT) database [19] and the 2016 Speakers In The Wild (SITW)[20] database. Our studies demonstrate the superiority of our proposed systems, especially with short utterances, and analyse the impact of speech utterance duration on speaker identification performance.

The system has been evaluated using AWGN, original speech recordings and various non-stationary noise (NSN) types, such as crowd chatter, street traffic and bus interior. All noise was added during the testing phase. However, the G.712 handset type at 16 kHz was used for both the training and testing phases. In this study, a wide range of UBM mixture sizes and SNR levels are used. In the final analysis, the system presents fair comparisons with other state-of-the-art techniques as well as the corresponding fusion-based GMM-UBM system.

This paper is organized as follows. The introduction is given in section 1. Section 2 overviews earlier research in speaker recognition with short utterances and the application of DNN techniques in the speaker recognition domain. The proposed approach for speaker identification, experimental setup, and database are provided in Section 3 and results are discussed in Section 4.

II. Related Work

These days, more and more studies are using speaker recognition in various applications. To carry out tasks properly, users must authenticate their identity due to the increasing number of voice-controlled interactive devices like smartphone assistants, the Internet, and remote navigation and control systems. Furthermore, given the exponential growth of information and multimedia data in recent decades, integrating speaker recognition into multimedia data indexing is essential for minimizing document research [1-2]. However, in many circumstances, collecting an adequate amount of speech data is more challenging.

Forensic applications and surroundings can result in voice samples that are fragmented, unclear, recorded in noisy environments, or interrupted with little speech content. Furthermore, users could be more reluctant to provide a lot of speech data, especially during test phases such as phone banking applications. Collecting information can be significantly more difficult by additional elements such as the speaker's personal characteristics or health status. Furthermore, practical applications could impose restrictions on the system, like memory and computing resource limits or fixed utterance duration requirements [1]. These factors make it more challenging to gather the enormous amount of data that conventional speaker recognition methods usually require. As a result, in order to effectively deal with these issues, research on the recognition of short has become essential.

Many techniques have been explored and proposed for speaker recognition in the last few decades. Significant progress has been made in automating voice-based identification of individuals, with the most effective state-of-the-art applications using GMM-Universal Background Models (GMM-UBM) and Gaussian Mixture Models (GMM). In telephone speech with a population of 49 speakers, these models have demonstrated their ability to produce robust speaker identifications, achieving high recognition accuracy of more than 90% with clean audio data and more than 80% with 30 seconds training time [22]. In addition, supervectors and Support Vector Machines (SVM) have become popular techniques for accurately analysing speakers. Recently, i-vector models and joint factor analysis have also been explored [23].

While speaker recognition studies have produced excellent results with vast volumes of speech data, these state-of-the-art techniques become much less efficient when dealing with fewer speech samples (less than 15 seconds) [24]. Research has demonstrated that decreasing the test speech duration from 20 seconds to 2 seconds can significantly increase the Equal Error Rate (EER), with values on NIST SRE databases rising from 6.34% to 23.89%. Moreover, as recent studies have shown, the EER can increase to as high as 35.00% when test speech durations fall below 2 seconds [25]. For several years, researchers have been focused on the study of short utterance speaker recognition. For instance, achieving adequate results in short utterances with the GMM [22] in speaker recognition technique is challenging, which uses segmented statistical features of the speech spectrum.

Despite this, state-of-the-art methods demonstrate noticeable constraints and significant performance degradation when dealing with short-term speech in spectral statistics. A study found that recognition performance reached 96% with 10 seconds of test speech duration but decreased to 79% with only 3 seconds. The training phase included a large amount of speech data lasting 100 seconds, indicating the significant challenge posed by short utterances, particularly in limited speaker samples [13]. Many efforts have been made to address short utterance challenges, and researchers have explored various automatic speaker recognition techniques, like feature extraction and modelling techniques, score normalization, and phonetic information integration techniques [14].

Furthermore, there is a significant focus on i-vector and probabilistic linear discriminant analysis (PLDA) based speaker recognition systems, regarded as state-of-the-art due to their ability to improve performance while accommodating shorter utterance durations. Studies have shown that employing proposed systems over 10% improves the Equal Error Rate (EER) as compared to baseline systems [12].

For the Speaker Identification System (SIS), 400 speakers were used. With an i-vector approach and without fusion, this technique achieved 39.3% and 49.5% SIAs under Original Speech Recording (OSR) and white noise, respectively, at a signal-to-noise ratio (SNR) of 15dB[26]. On the other hand, under identical conditions, the same system employing the GMM-UBM without fusion approach achieved 39.7%, 24.6% (15dB) respectively.

In addition, ZT-normalization was also used to enhance the GMM-UBM system, achieving 42.5% and 29.7% (15dB) .50 speakers were chosen using an MIT mobile phone, and a corpus was developed. Afterwards, 94.14% and 92.36% of the SIAs under OSR at the Cosine Distance Scoring (CDS) and Support-Vector Machines (SVM) were achieved, respectively, using the i-vector without fusion approach in conjunction with Linear Discriminant Analysis (LDA) and Within-Class Covariance Normalisation (WCCN) [27]. Furthermore, the retrieval i-vector without fusion approach was used with 1,000 speakers from YouTube [4].

Using a testing duration of 10 seconds, the system yields 92% SIA under OSR. This has been the first study to use the i-vector technique for speaker identification, utilising the 2016 SITW database. As a result, the SITW, NIST 2008, and TIMIT datasets were the focus of this paper's comprehensive analysis of the coupling of the ELM-i-vector technique with different noise and handset effects. In addition, the study compares the performance of the i-vector with the GMM-UBM technique [28] (examined in the preceding section) using the same environment and database. Thus, this work establishes a standard for future speaker identification studies. The system was evaluated using the TIMIT database and employed the i-vector with the PLDA, attaining SIAs of 95% for long speech (6-9 s) and 85% for short speech (2-4 s)[29].

Using the TIMIT database in the same setting, our system, which used the i-vector-ELM technique for speech length 8s, achieved 96.67%, which was used in [28] and outperformed the work [29].

With the advancement of deep learning technology, deep neural networks (DNNs) are being employed widely in the field of speaker recognition. DNNs have demonstrated significant performance increases in various pattern

recognition applications, including voice and face recognition. DNNs have been successfully utilised as a substitute or to improve existing approaches, especially the Probabilistic Linear Discriminant Analysis (PLDA) approach based on i-vectors [4,12,27,29]. Researchers have shown favourable results when using DNN acoustic models instead of Gaussian mixture models to extract sufficient statistics for speaker recognition.

In one study, a proposed framework utilizing DNNs achieved a relative improvement of 30% in Equal Error Rate (EER) when evaluated on telephone conditions compared to a state-of-the-art system [30]. Additionally, using DNN bottleneck features instead of traditional Mel-Frequency Cepstrum Coefficients (MFCC) also improved performance. Research on end-to-end systems has also been conducted, and these systems have shown competitiveness for short test utterances and a large training data set for text-independent speaker recognition.

Furthermore, recurrent neural networks (RNNs) have demonstrated promising effectiveness in speaker recognition, especially when using Long Short-Term Memory (LSTM) architecture [16-17]. Many research investigations have additionally used the Convolutional Neural Network (CNN) based models for speaker recognition; specific architectures have proven more efficient than traditional methods like i-vector-based techniques [15]. Considering the significance of short utterance speaker identification in modern applications, we aim to examine how well DNN techniques perform when improving speaker identification systems under short utterance evaluation conditions [30].

We hypothesise that using CNN and LSTM networks in combination with RNN models can provide better adaptable classifiers capable of acquiring variations observed in short speech utterances since short utterances may contain fewer speaker features than longer ones. Therefore, in order to overcome the constraints caused by shorter data lifetime, we will suggest and assess two DNN system topologies that are strengthened with recently modified characteristics.

III. METHODOLOGY

This article presents the development of a speaker identification system that makes use of deep learning techniques to operate in constrained conditions with short data durations in order to achieve satisfactory performance.

The acoustic data is first preprocessed, and then the required features are extracted using Mel-Frequency Cepstral Coefficients (MFCC) [9,31-35]. In addition, we propose using Cepstral Mean and Normalisation or CMVNC features to offer improved signal and time variation adaptability. These coefficients increase the system's robustness against slowly fluctuating additive noise and linear channel effects by utilising sliding windows to normalise the distribution parameters of cepstral coefficients across predetermined time intervals. This enhancement aims to increase the system's capacity to extract maximum information from small voice data segments.

The suggested system uses CNN and LSTM-RNN classification models for speaker recognition tasks. Figure 1 shows the flow diagram of the first suggested system that uses CNN models, and Figure 2 shows the flow diagram of the second proposed system that uses LSTM-RNN models. We compute speaker accuracy and compare both the modals in order to verify the effectiveness of our approach.

A. Acostic Features

In order to obtain the most effective parametric representation of acoustic signal, feature extraction, an essential step in recognising speakers recognition systems, is extracted during the acoustic preprocessing phase. Mel-Frequency Cepstral Coefficients (MFCCs) [9,31] are considered the most effective among those used in speaker recognition because they closely resemble the way sounds are perceived by humans.

Various modern technologies in speech and speaker recognition have encouraged the use of MFCC features in combination with Deep Neural Networks (DNNs) [30]. In order to extract linguistic content while removing unnecessary information like noise and emotions, choosing relevant input data is essential to speaker recognition.

This work uses a 20 ms Hamming window with a 10 ms overlap to extract MFCC features. Every 10 ms, the first and second derivatives of the twelve MFCC coefficients are calculated, and the result is a 45-dimensional feature

vector for each frame. Recent research projects, as well as state-of-the-art speaker recognition systems, frequently use this feature vector structure.

We use Cepstral Mean Normalisation (CMN) to reduce the variability in the extracted characteristics between speech sessions. We also apply Cepstral Mean and Variance Normalisation (CMVN) to further improve robustness. Next, we employ the Mean and Variance Normalised MFCCs, or CMVNC coefficients, which are used to present a short-time cepstral representation of speech signal. The CMVN approach is used to normalise feature vector coefficients.

In particular, the coefficient of a given feature vector, the resulting feature vector representing the MVNMFCC [32] coefficients, is computed as follows for a given feature vector X = {x[1], x[2],..., x[N]} of MFCC coefficients:

$$\hat{x}[n] = \frac{x[n]-\overline{x}}{\sigma_x} \quad (1)$$

$$\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x[n] \quad (2)$$

$$\sigma_x^2 = \frac{1}{N}\sum_{n=1}^{N} (x[n]-\overline{x})^2 \quad (3)$$

In this case, n denotes the MFCC coefficient's order within the feature vector, and N indicates the total number of MFCC coefficients in the feature vector.

B. Modelling Approaches

1) Convolutional Neural Network (CNN): Now days Deep Neural Network (DNN) architectures are used in several applications like speech recognition, machine translation, computer vision, natural language processing and bioinformatics [1,2,3,5]. DNNs [30] allow complex data to be modelled with fewer layers but more expressive features by adding more hidden layers between the input and output layers. The researchers have given Convolutional Neural Networks (CNNs) a lot of attention, and these networks are now becoming the main research tool in image and speech processing. The Convolutional Neural Networks (CNNs) is a type of Deep Neural Network that simulates how the brain's visual cortex processes and analyses images. Its main objective is to find local structures in the incoming data [15].

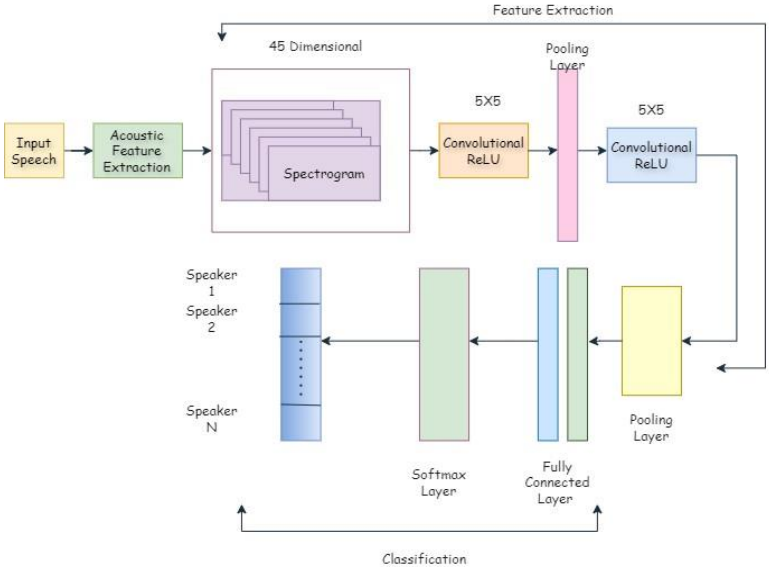

Fig.1: Proposed CNN Architecture for Speaker Recognition

Spectrograms [9] represent variations in the signal spectrum and offer extensive information about a speaker's unique features in the field of speaker recognition. As a result, feature vectors derived from spectrograms are considered adequate when CNNs utilised in signal-based applications.

Speech is a signal that varies over time and has complex relationships across a range of time periods; Spectrograms offer reliable tools for observing variations in the speech signal. As such, CNNs use spectrograms as their input, which allows for transformation invariance in both space and time, maintaining the temporal structure of the voiceprint information. Therefore, spectrograms are recommended as CNN inputs by the speaker recognition researchers. The CNN model consists of convolutional layers, pooling layers, fully connected layers, activation functions, and a classification stage, including a dropout layer that enables regularisation to reduce overfitting [15].

Speech signals are usually preprocessed using the Short Time Fourier Transform (STFT) on speech segments to produce spectrograms suitable for CNN classification. This study uses MFCC [9,31] and MVNMFCC features to more accurately represent speech information in the dataset by using essential acoustic properties.

In particular, speech signals are split up into frames, and then for each frame a 20 ms window size is applied and a 10 ms overlap of Hamming windows. After extracting 45cepstral features from every frame, STFT is used to create spectrograms for every frame. Spectrograms are used to represent individual voice signals.

The proposed CNN model in Fig.1 consists of three convolutional layers and three max pool layers. For the input layer, frames of 32-dimensional filter-bank features are grouped together to generate a feature map. The kernel size of each convolutional layer is 5 x 5, and its stride is 2 × 2. During the training phase, activation is done using the ReLU activation function [15], which is well-known for its efficiency in deep learning.

Because it is sensitive to internal covariate changes, each convolutional layer is connected to a 3 × 3 max pooling layers prior to Batch Normalisation is applied to improve speaker representation, followed by flattened and dense layers. The dense layer is used for classification, which has a size of 32, considering the 32 classes, and it utilises the SoftMax function [30] for scoring.

2) Recurrent Neural Network: Recently, research has emphasised the remarkable performance of Recurrent Neural Networks (RNNs) in language modelling tasks. RNN models are popular in many applications and are especially well-suited for sequential data. Sequence-to-sequence models have become popular for tasks like speech identification because of the development of deep learning [16-17]. RNNs are not like typical neural networks in that they may make decisions based on inputs at that moment and use information from past occurrences to predict future ones. In simple terms, the input at a given time step and the output from previous time steps determine the output at that particular time step. RNN's memory of computed outputs is stored by each cell, affecting decisions at subsequent steps. This feature shows that previous outputs, as well as the current input, have an influence on how the network reacts to new data.

Mathematically, for a given input time series $x = \{x_1, x_2, ..., x_T\}$, the RNN calculates the output sequence $y = \{y_1, y_2, ..., y\_T\}$ and hidden state sequence $h = \{h_1, h_2, ..., h_T\}$ iteratively by using the equations 4 and 5:

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \qquad (4)$$

$$y_t = g(W_{hy}h_t + b_y) \quad (5)$$

where W is weight matrices, b_h and b_y denote bias for the hidden and output layers, respectively, and f(.) and g(.) are activation functions for the hidden and output layers.

RNNs can extract information from previous steps using the hidden state h_t at time t still experience problems like the vanishing and exploding gradient during back propagation, which can lead to a decrease in performance.

Because of this drawback, RNNs cannot effectively manage long-term dependencies in sequential data since they may find it difficult to retain information from earlier time steps in longer sequences. RNNs are essentially short-term memory systems that can forget essential details from longer sequences.

3) Long Short-Term Memory (LSTM) The inherent limitation of the traditional Recurrent Neural Network (RNN) models, which have trouble retaining information over long sequences, was addressed with the introduction of Long Short-Term Memory (LSTM) networks [16-17]. LSTMs are specialised RNN architectures with memory cells designed to help with this problem by making it easier to detect long-term dependencies. Due to their unique memory cell layout, long-range dependency problems benefit significantly from LSTMs' ability to efficiently reduce the vanishing gradient problem.

In the context of an input time series $x = \{x_1, x_2,..., x_T\}$, LSTM networks iteratively map the input time series to two output sequences: $h = \{h_1, h_2,..., h_T\}$ and $y = \{y_1, y_2,..., y_T\}$ by updating the states of memory cells. This process involves several steps, defined by the following equations:

$$f(t) = \sigma(w_{fx}x_t + w_{fh}h_t + w_{fc}c_{t-1} + b_f) \quad (6)$$

$$i(t) = tan(w_{ix}x_t + w_{ih}h_{t-1} + w_{ic}c_{t-1} + b_i) \quad (7)$$

$$U(t) = tanh(w_{cx}x_t + w_{ch}h_{t-1} + b_c) \quad (8)$$

$$C(t) = (U_t i_t + C_{t-1}f_t) \quad (9)$$

$$o(t) = tanh(w_{ox}x_t + w_{oh}h_{t-1} + w_{oc}c_{t-1} + b_o) \quad (10)$$

$$h(t) = (O_t * tanh\ (C_t) \quad (11)$$

$$y_t = k(W_{yh}h_t + b_y) \quad (12)$$

Here, the input weight matrices are denoted by W_{ix}, W_{fx}, W_{ox}, and W_{cx},, while the recurrent matrices are represented by W_{ih}, W_{fh}, W_{oh}, W_{ch}. W_{ix}, W_{ix}. The hidden output weight matrix is represented by W_{yh} , and W_{ic}, W_{fc}, W_{oc} represent the weight matrices of peephole connections. The corresponding bias vectors are given by b_i, b_f, b_o, b_c, b_y. Logistic sigmoid function is , i,f,o and c represent the input gate, forget gate, output gate, and cell activation vectors, respectively. The hyperbolic tangent function is tanh, serving as the activation function for the output network. In our experiments, we used the ReLU activation function. We employ an LSTM RNN network-based speaker recognition system in this work. Our proposed system structure is shown in Fig.2.

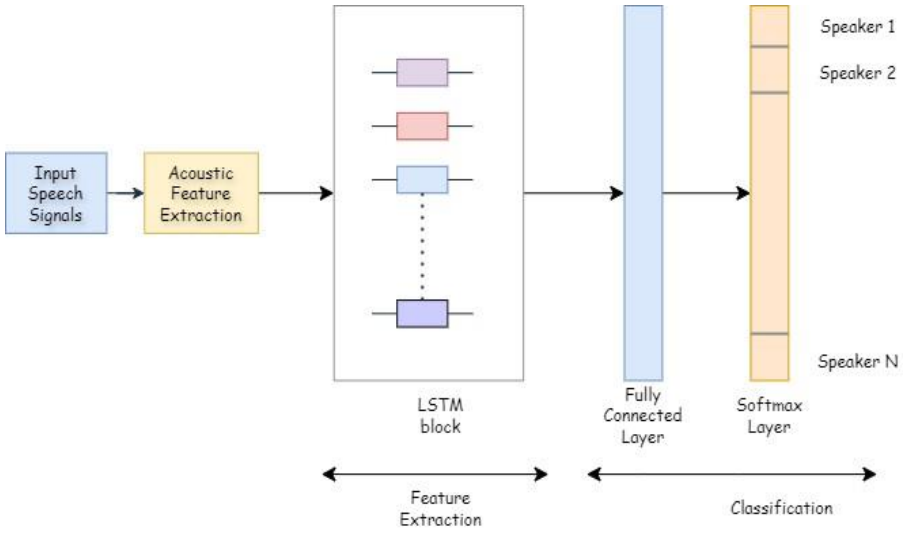

Fig.2: The Proposed LSTM-RNN-based recognition system

The proposed approach given in Fig.2 is used to evaluate performance in terms of Precision, Recall and accuracy for training duration of 10s and 20s. The system is for different testing duration for datasets 2008NIST, TIMIT and SITW 2016.

C. Databases and Environments

This research evaluates the three main databases—TIMIT, SITW 2016, and 2008 NIST. According to dialect regions, 120 speakers from the TIMIT database were chosen for this study. For 1200 voice utterances, a fixed speech length is obtained (for training 720 and for testing 480 utterances). However, AWGN and Non-Stationary Noise (NSN) were included during the testing phase.

As in [32], seven SNR levels are generated (0 dB to 30 dB) based on the corresponding noise power, with a set step size (5 dB) for each level. The NSN, which was used in the testing phase, was previously described [4]. The AWGN and NSN noise files have been reduced to match the fixed length of the original voice samples, which was 129,250 (8 s). Furthermore, a G.712 model cellphone is utilised for training and testing concerning the normalised speech phases.

The 4^{th}-order linear Infinite Impulse Response (IIR) filter in the G.712 type phone is utilized [32]. However, the Speakers in The Wild (SITW) 2016 database has been collected from open-source media under different challenging situations like stadiums, outdoor settings, and red-carpet interviews with one or more speakers. This database's primary goal is to benchmark speech identification technologies by encouraging researchers to create innovative algorithms.

120 speakers were used in this work; those were chosen from single and unbalanced multi-speaker sets. Furthermore, this work used audio edition tools, specifically Audacity and Gold Wave, to determine the target speaker while the interviewer is ignored and used to produce a single speaker. Ten equal lengths have been obtained by dividing each speech file, with 129,250 samples fixed length (which comes from the multiplication of the sampling frequency 16 kHz with the speech length 8.078125 s (about 8 s) to yield 129,250 samples), in order to match other databases used in this work as well as in previous work [9].

Furthermore, in order to achieve a constant length, concatenation techniques are applied to speech files that are shorter than 8 seconds. Four files were used for testing and six for training. However, the NIST 2008 dataset comprises telephone speech recordings and multilingual microphone recordings of native and bilingual English interview speakers.

To be comparable to other databases employed in this research, the original speech files, which had an 8 kHz sampling frequency, were changed to a 16 kHz sampling frequency [4]. A sample of 120 English speakers was chosen from the microphone channel for comparison with other databases. Again, after deleting the interviewers, we were left with only single speakers. We next generated four testing and six training speech recordings, each of which had a fixed duration of eight seconds.

IV. Experimental Results

To determine its effectiveness, we evaluated our speaker recognition system on three commonly used datasets in this field [5, 38]. We used speech samples from the 2008 NIST, the SITW 2016, and the TIMIT databases.

We performed the experimental evaluations mentioned below to evaluate the effectiveness of our proposed speaker identification system and enable comparison with prior studies [4,12,27]. We introduce two proposed systems, one based on the LSTM-RNN technique and the other on CNN models. These systems are implemented and evaluated across various conditions.

We take into account different durations for training and testing speech utterances in order to thoroughly evaluate their performances. To be more precise, we tested using speech segments of 1.5, 1 and 0.5 seconds per speaker, and we evaluated using utterances of 20 and 10 seconds per speaker for the training.

We first perform experiments using MFCC features, then switch to CMVNC acoustic features for further comprehensive analysis. We aim to explain the contributions of our suggested technique to speaker identification with short utterances by comparing results from different approaches.

A. Speaker Recognition with 20s of Training Data Duration

We start our research by comparing how different speech recognition systems perform in different experiment conditions. We follow up with two proposed systems, one based on the LSTM-RNN technique and the other on CNN models. MFCC coefficients are evaluated as acoustic features for these systems.

Three convolutional layers with a kernel size of (5,5) and Rectified Linear Unit (ReLU) activation functions are used in the CNN model. Each convolutional layer is connected to a MaxPooling layer with a pool size of (3,3) after batch normalisation. After applying a flatten layer, the outputs are fed into a dense layer with the same ReLU activation function. A dropout rate of 20% has been added to avoid overfitting. Finally, the number of speakers in the dataset is represented by using a fully connected layer matching with 32 classes and using the SoftMax activation function.

The LSTM-RNN system consists of an input layer, three hidden layers and a dense layer with ReLU activation function. A dropout rate of 20% is used to reduce overfitting. A fully connected layer with SoftMax activation is utilised, corresponding to the number of speaker classes (16-17).

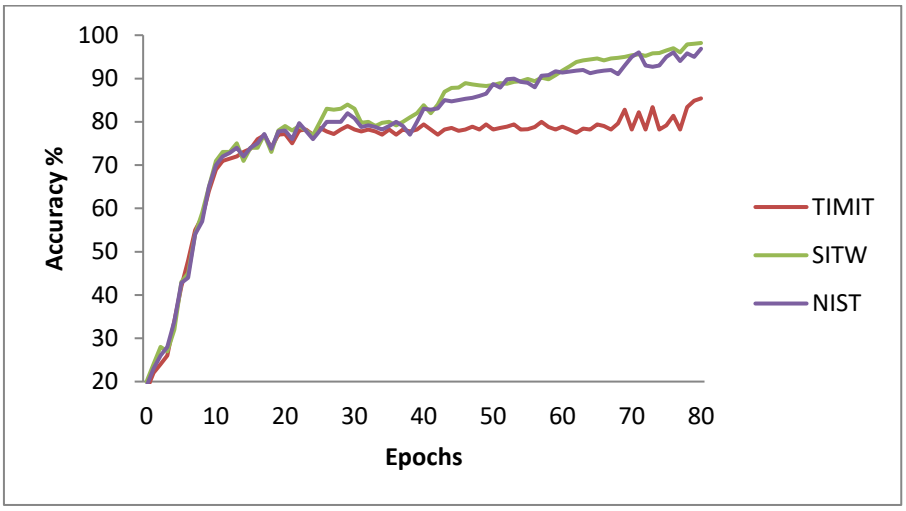

Fig.3: Epoch Vs. Accuracy curve for 10s training and 1.5s testing data using proposed LSTM-RNN for all the databases.

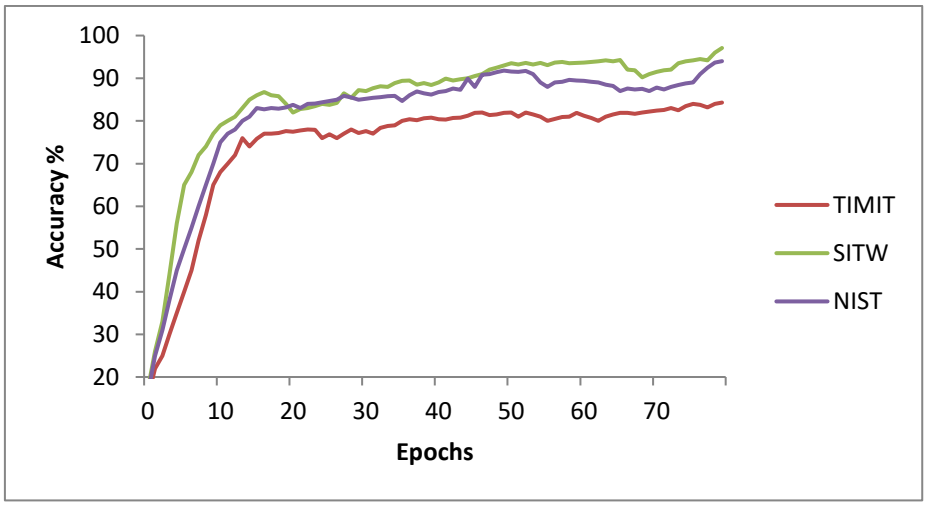

Fig.4: Epoch Vs. Accuracy curve for 10s training and 1.5s testing data using proposed CNN for all the databases.

The CNN and LSTM-RNN models perform training across 80 epochs, with accuracy provided as the performance metric in Fig.4. Plotting accuracy curves show how the system performs under various experimental conditions. Fig. 5 is the plot of Accuracy vs Epochs with all three datasets with the CNN approach. Accuracy increases as the model is trained for more epochs.

Shorter test speech utterances (0.5 s) result in decreased performance, as evaluated on the TIMIT dataset; the CNN and LSTM-RNN systems only achieve accuracies of 72.45% and 76.24%, respectively, as given in Table 1.

Similar trends occur when the SITW 2016 dataset is used to evaluate the systems. For example, the CNN obtains an accuracy of approximately 95.18 % with 1.5-second test utterances, which is higher than the LSTM RNN-based system when evaluated with 0.5-second utterances; however, the LSTM-RNN system performs better than CNN-based systems, with accuracies of 96.87% as compared to CNN 93.57% for NIST 2008 database, with 1.5s testing utterances as in Table 2.The LSTM-RNN system outperforms the NIST database's CNN systems, achieving accuracy of 86.76% and 88.38% with 1-second utterances, as given in Table 3.

To improve the system performance, Normalising cepstral coefficients and CMVNC features are used, which more accurately capture spectrum variations in speech signals

Precision and recall metrics are computed to provide further insights into system performance. Precision, indicating the ratio of true predicted positives to total predicted positives, ranges from 96.28% to 88.8%, while recall, reflecting the ratio of correctly predicted positive observations, ranges from 94.6% to 92.19% for LSTM-RNN for 1.5s testing duration. Our proposed system demonstrates competence in short utterance speaker recognition, showcasing its effectiveness in all three datasets.

Table 1 Speaker recognition accuracy for 10s and 20s Training and different testing durations with the proposed CNN-based and LSTM RNN system for the TIMIT database

Test duration	Proposed Approach	10s Training			20s Training		
		Precision	Recall	Accuracy	Precision	Recall	Accuracy
1.5s	CNN	95.20	94.50	82.28	96.9	95.5	83.9
1s		90.81	87.88	74.53	93.5	88.1	76.45
.5s		80.87	75.67	71.23	80.95	75.9	72.45
1.5s	LSTM-RNN	96.28	94.6	84.3	97.28	95.8	85.4
1s		92.39	90.49	75.03	94.8	89.4	77.24
.5s		81.02	78	75.55	83.45	76.7	76.24

Table 2 Speaker recognition accuracy for 10s and 20s Training and different testing duration with the proposed CNN-based and LSTM-RNN system for SITW 2016 database

Test duration	Proposed Approach	10s Training			20s Training		
		Precision	Recall	Accuracy	Precision	Recall	Accuracy
1.5s	CNN	89.98	92.59	94.07	90.2	93.6	94.81
1s		82.59	90.20	82.58	84.5	92.5	84.24
.5s		75.59	88.59	71.18	76.45	89.2	72.20
1.5s	LSTM-RNN	88.8	92.80	95.09	91.4	94	96.47
1s		86.78	93.54	86.9	87.5	92.9	89.7
.5s		77.60	91.55	78.54	77.52	90.2	80.24

Table 3 Speaker recognition accuracy for 10s and 20s Training and different testing duration with the proposed CNN-based and LSTM-RNN system for the NIST database

Test duration	Proposed Approach	10s Training			20s Training		
		Precision	Recall	Accuracy	Precision	Recall	Accuracy

1.5s	CNN	91.18	93.50	91.15	91.4	93.82	93.57
1s		85.59	91.18	85.54	87.28	92.67	86.76
.5s		82.26	90.28	82.08	83.9	92.59	84.59
1.5s	LSTM-RNN	94	92.19	94	95.4	91.27	95.24
1s		85.59	93.33	86.08	86.5	93.94	88.38
.5s		88.23	95.82	88.52	89.9	94.37	90.98

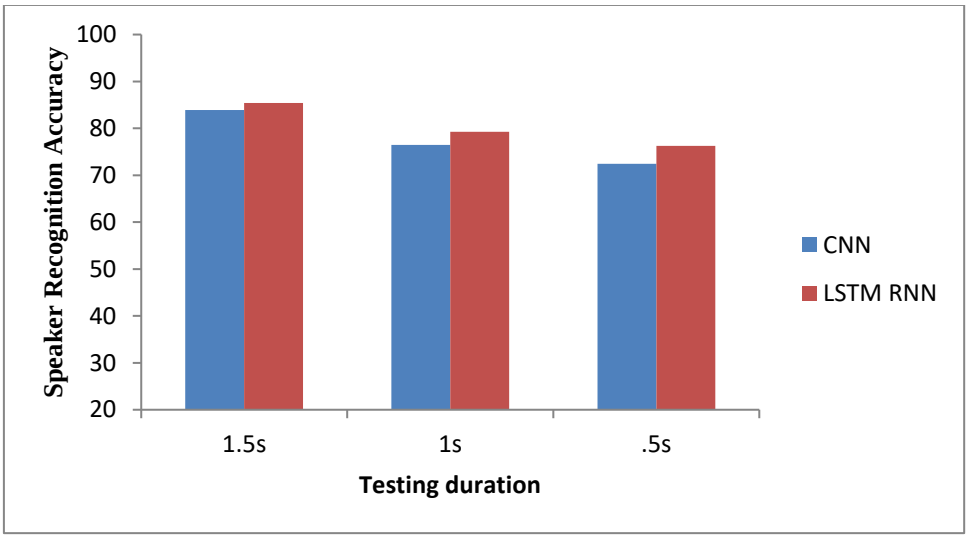

Fig. 5: Speaker recognition performance with CNN & proposed LSTM-RNN with TIMIT database with 20s training time

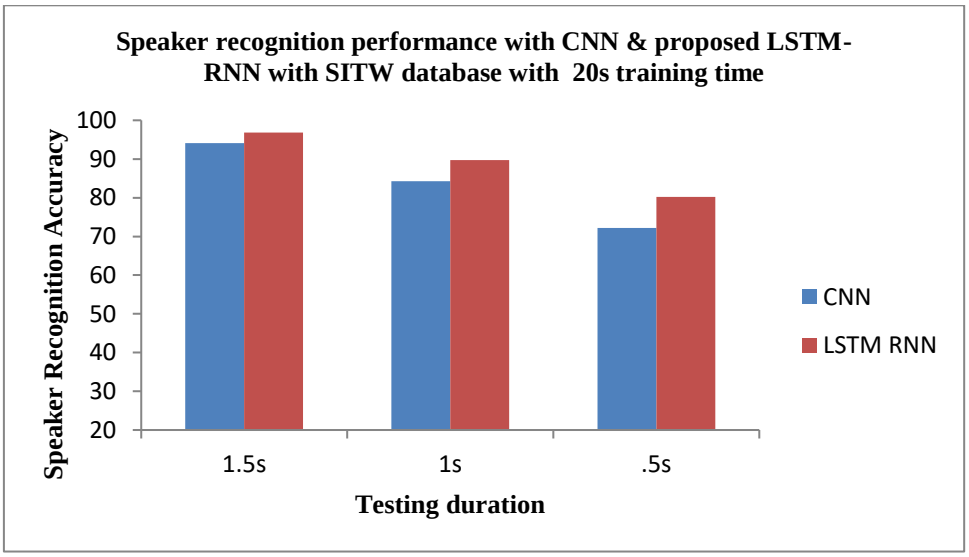

Fig. 6: Speaker recognition performance with CNN & proposed LSTM-RNN with SITW 2016 database with 20s training time

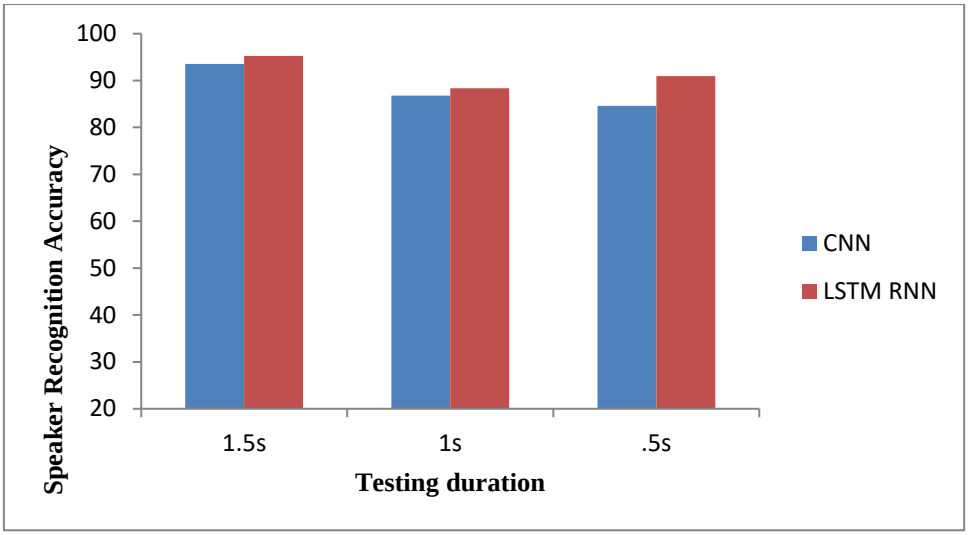

Fig.7: Speaker recognition performance with CNN & proposed LSTM-RNN with NIST database with 20s training time

The accuracy of the proposed LSTM-RNN and CNN is 79.24% and 76.45% with the TIMIT database,1s testing time and 20s training time as given in Fig. 5. We obtained maximum accuracy as in Fig. 6 96.87% with SITW 2016 database and 94.07% by CNN modal with 1.5s test time. Recognition performance achieved 93.57% by CNN modal with 1.5s test duration dropped to 84.59% for .5s test duration as in Fig.7.

B. Speaker Recognition with 10 s of Training Data Duration

This section assesses the speaker recognition performance of various baseline and proposed systems when trained with limited data duration. Experiments are conducted with only 10 seconds of training data, following the same setup for the test task previously described.

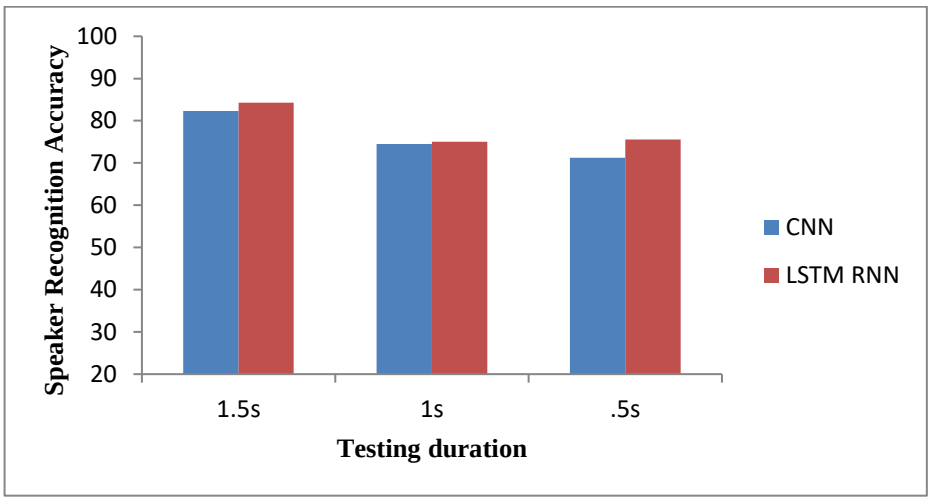

Fig. 8: Speaker recognition performance with CNN & proposed LSTM-RNN with TIMIT database with 10s training time

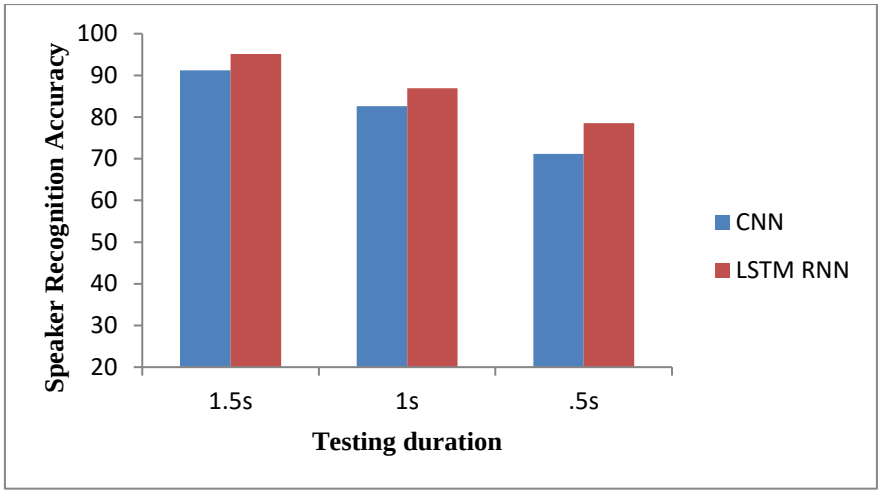

Fig. 9: Speaker recognition performance with CNN & proposed LSTM-RNN with SITW 2016 database with 10s training time

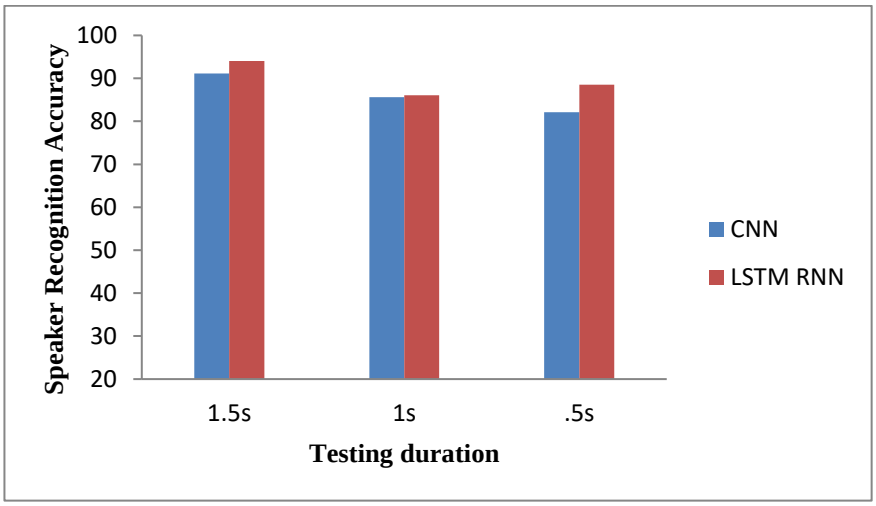

Fig. 10: Speaker recognition performance with CNN & proposed LSTM-RNN with NIST database with 10s training time

Fig. 8 illustrates the speaker recognition accuracy obtained from the CNN and LSTM approach, where 10 seconds of speech utterances per speaker are used for training, and segments of lengths 1.5, 1, and 0.5 seconds are used for testing across TIMIT database. The results depict the performance of the LSTM-RNN and CNN-based systems using CMVNC coefficients for TIMIT dataset.

Fig. 9 and Fig. 10 show the best speaker recognition accuracy obtained from the LSTM RNN modal with a testing duration of 1.5s, which is used in the testing for all the different databases.

In general, we observe that performance tends to decrease as the duration of speech data decreases. However, when employing CMVNC coefficients, the proposed LSTM-RNN system consistently outperforms both the i-vector-PLDA and CNN-based systems. For example, with the TIMIT database, the proposed LSTM-RNN system achieves a recognition accuracy of 84.3% with 1.5 seconds of test utterances, surpassing the CNN-based system (82.28%). Even with shorter test durations of 1 and 0.5 seconds, the proposed LSTM-RNN system performs competitively, outperforming both baseline systems.

Similar trends are observed with the SITW 2016 and NIST 2008 databases, where the proposed LSTM-RNN system consistently achieves higher recognition accuracy compared to CNN-based systems across different test durations and 10s training time as shown in Fig. 8 to Fig. 10. As the training time is doubled, i.e. 20s duration there is a significant improvement in accuracy with different testing time. There is an increase in accuracy of

85.4%, 96.47% and 95.24% for datasets TIMIT, SITW 2016 and NIST 2008, respectively, as depicted in Fig.5, Fig.6 and Fig.7.

The evaluation results suggest that the proposed LSTM-RNN system using CMVNC coefficients significantly improves speaker recognition performance compared to baseline systems, even when trained with limited data duration. This approach represents a valuable solution for enhancing the performance of speaker recognition systems, mainly when dealing with reduced speech data duration in both the training and testing phases.

To provide a clearer comparison of the proposed systems' contributions summarizes the recognition accuracy of the proposed LSTM-RNN system using CMVNC coefficients and the CNN-based system using CMVNC coefficients against those using MFCC coefficients. These comparisons underscore the effectiveness of the proposed approach in improving speaker recognition performance across various experimental setups and datasets.

V. CONCLUSION

This study introduces and assesses a novel speaker recognition system based on deep learning methodologies. We propose two distinct architectures: one utilizing a convolutional neural network (CNN) with CMVNC coefficients and another employing a long short-term memory recurrent neural network (LSTM-RNN) with the same coefficients. Our proposed systems outperform the baseline i-vector-PLDA system and CNN and LSTM-RNN approaches using MFCC coefficients. We conduct evaluations across various training and testing data durations using the NIST 2008, SITW 2016, and TIMIT 2008 databases under specific conditions. The results demonstrate the effectiveness of our novel approach in improving recognition performance, particularly in addressing the challenges posed by short utterances in speaker recognition tasks. However, our system exhibits limitations with extremely short utterance durations, suggesting opportunities for further enhancement. Future research will focus on integrating additional modelling techniques to augment the proposed system's performance, explicitly addressing more challenging scenarios. We also plan to explore integrating face detection techniques alongside audio signal processing to enhance recognition accuracy through multimodal approaches.

REFERENCES

[1] Furui, Sadaoki. "An overview of speaker recognition technology." Automatic Speech and Speaker Recognition: Advanced Topics (1996): 31-56.

[2] Hanifa, RafizahMohd, Khalid Isa, and Shamsul Mohamad. "A review on speaker recognition: Technology and challenges." Computers & Electrical Engineering 90 (2021): 107005.

[3] Beigi, Homayoon, and HomayoonBeigi. Speaker recognition. Springer US, 2011

[4] Verma, Pulkit, and Pradip K. Das. "i-Vectors in speech processing applications: a survey." International Journal of Speech Technology 18 (2015): 529-546.

[5] Singh, Maneet, Richa Singh, and Arun Ross. "A comprehensive overview of biometric fusion." Information Fusion 52 (2019): 187-205.

[6] Yasmin, Ghazaala, Subrata Dhara, RudrenduMahindar, and Asit Kumar Das. "Speaker identification from mixture of speech and non-speech audio signal." In Soft Computing in Data Analytics: Proceedings of International Conference on SCDA 2018, pp. 473-482. Springer Singapore, 2019.

[7] Pawar, M. D., and Rajendra Kokate. "A robust wavelet based decomposition and multilayer neural network for speaker identification." In Innovations in Electronics and Communication Engineering: Proceedings of the 7th ICIECE 2018, pp. 197-209. Springer Singapore, 2019.

[8] Modak, Sandip Kumar Singh, and Vijay Kumar Jha. "Multibiometric fusion strategy and its applications: A review." Information Fusion 49 (2019): 174-204.

[9] Kumari, R. Shantha Selva, and S. Selva Nidhyananthan. "Fused MEL feature sets based text-independent speaker identification using Gaussian mixture model." Procedia Engineering 30 (2012): 319-326.

[10] Lai, Jun-Yao, Shi-Lin Wang, Alan Wee-Chung Liew, and Xing-Jian Shi. "Visual speaker identification and authentication by joint spatiotemporal sparse coding and hierarchical pooling." Information Sciences 373 (2016): 219-232.

[11] M.A. Islam, W.A. Jassim, N.S. Cheok, M.S.A. Zilany, A robust speaker identification system using the responses from a model of the auditory periphery. PLoS ONE 11(7), 1–21 (2016)

[12] Kanagasundaram, Ahilan, David Dean, and Sridha Sridharan. "Improving PLDA speaker verification with limited development data." In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1665-1669. IEEE, 2014.

[13] Li, K. P., and E. H. Wrench Jr. "Text-independent speaker recognition with short utterances." The Journal of the Acoustical Society of America 72, no. S1 (1982): S29-S30.
[14] Fatima, Nakhat, and Thomas Fang Zheng. "Short utterance speaker recognition a research agenda." In 2012 international conference on systems and informatics (ICSAI2012), pp. 1746-1750. IEEE, 2012.
[15] Hourri, Soufiane, Nikola S. Nikolov, and Jamal Kharroubi. "Convolutional neural network vectors for speaker recognition." International Journal of Speech Technology 24, no. 2 (2021): 389-400.
[16] El-Moneim, Samia Abd, M. A. Nassar, Moawad I. Dessouky, Nabil A. Ismail, Adel S. El-Fishawy, and Fathi E. Abd El-Samie. "Text-independent speaker recognition using LSTM-RNN and speech enhancement." Multimedia Tools and Applications 79 (2020): 24013-24028.
[17] Dua, Mohit, Pawandeep Singh Sethi, Vinam Agrawal, and Raghav Chawla. "Speaker recognition using noise robust features and LSTM-RNN." In Progress in Advanced Computing and Intelligent Engineering: Proceedings of ICACIE 2020, pp. 19-28. Springer Singapore, 2021.
[18] Mikaia, Anzor, Principal Edward White V. EI, Vladimir Zaikin EI, Damo Zhu EI, O. David Sparkman EI, Pedatsur Neta, Igor Zenkevich RI et al. "NIST standard reference database 1A." Standard Reference Data, NIST, Gaithersburg, MD, USA https://www. nist. gov/srd/nist-standard-reference-database-1a (2014).
[19] Lopes, C., &Perdigao, F. (2011). Phone recognition on the TIMIT database. Speech Technologies/Book, 1, 285-302
[20] McLaren, M., Ferrer, L., Castan, D., & Lawson, A. (2016, September). The speakers in the wild (SITW) speaker recognition database. In Interspeech (pp. 818-822).
[21] Campbell, Joseph P., Wade Shen, William M. Campbell, Reva Schwartz, Jean-Francois Bonastre, and DrissMatrouf. "Forensic speaker recognition." IEEE Signal Processing Magazine 26, no. 2 (2009): 95-103.
[22] McLaughlin, Jack, Douglas A. Reynolds, and Terry Gleason. "A study of computation speed-ups of the GMM-UBM speaker recognition system." In Sixth European conference on speech communication and technology. 1999.
[23] Dehak, Najim, Patrick Kenny, Reda Dehak, OndrejGlembek, Pierre Dumouchel, Lukas Burget, ValiantsinaHubeika, and Fabio Castaldo. "Support vector machines and joint factor analysis for speaker verification." In 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4237-4240. IEEE, 2009.
[24] Jayanna, H. S., and SR Mahadeva Prasanna. "Multiple frame size and rate analysis for speaker recognition under limited data condition." IET Signal processing 3, no. 3 (2009): 189-204.
[25] Mak, Man-Wai, Roger Hsiao, and Brian Mak. "A comparison of various adaptation methods for speaker verification with limited enrollment data." In 2006 IEEE international conference on acoustics speech and signal processing proceedings, vol. 1, pp. I-I. IEEE, 2006.
[26] Karadaghi, Rawande, Heinz Hertlein, and Aladdin Ariyaeeinia. "Effectiveness in open-set speaker identification." In 2014 International Carnahan Conference on Security Technology (ICCST), pp. 1-6. IEEE, 2014.
[27] Liu, Tingting, Kai Kang, and Shengxiao Guan. "I-vector based text-independent speaker identification." In Proceeding of the 11th World Congress on Intelligent Control and Automation, pp. 5420-5425. IEEE, 2014.
[28] S. Al-Kaltakchi, Musab T., Wai L. Woo, SatnamDlay, and Jonathon A. Chambers. "Evaluation of a speaker identification system with and without fusion using three databases in the presence of noise and handset effects." EURASIP Journal on Advances in Signal Processing 2017 (2017): 1-17.
[29] Nayana, P. K., Dominic Mathew, and Abraham Thomas. "Comparison of text independent speaker identification systems using GMM and i-vector methods." Procedia computer science 115 (2017): 47-54.
[30] Rohdin, Johan, Anna Silnova, Mireia Diez, OldřichPlchot, Pavel Matějka, LukášBurget, and OndřejGlembek. "End-to-end DNN based text-independent speaker recognition for long and short utterances." Computer Speech & Language 59 (2020): 22-35
[31] Gupta, Shikha, Jafreezal Jaafar, WF Wan Ahmad, and Arpit Bansal. "Feature extraction using MFCC." Signal & Image Processing: An International Journal 4, no. 4 (2013): 101-108.
[32] Al-Kaltakchi, Musab TS, Wai Lok Woo, Satnam Singh Dlay, and Jonathon A. Chambers. "Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification." In 2016 4th international conference on biometrics and forensics (IWBF), pp. 1-6. IEEE, 2016.
[33] Sharma, Geetanjali, Amit M. Joshi, and Emmanuel S. Pilli. "DepML: An efficient machine learning based MDD detection system in IoMT framework." SN Computer Science 3, no. 5 (2022): 394.
[34] Sethi, Manan, Karna Sharma, PaanshulDobriyal, Navya Rajput, and Geetanjali Sharma. "A Novel High Performance Dual Threshold Voltage Domino Logic Employing Stacked Transistors." International Journal of Computer Applications 77, no. 5 (2013).
[35] Sharma, Geetanjali, and Amit M. Joshi. "Novel eeg based schizophrenia detection with iomt framework for smart healthcare." arXiv preprint arXiv:2111.11298 (2021).