Available at https://www.iasj.net/iasj
Iraqi Academic Scientific Journals
Journal homepage:
https://journals.uokerbala.edu.iq/index.php/UOKJ

Research Article

Toward Robust Speaker Identification at Extreme Noise Levels: A CNN–BiLSTM Approach

[1,]Kawther Meitham Mohammed Jawad [2,] Ashwan A. Abdulmunem

[1,] Department of Computer Science, College of Computer Science and Information Technology, University of Karbala, Karbala, Iraq

Article Info
Article history:
Received 2 -3-2026
Received in revised form 18-5-2026
Accepted 14-6-2026
Available online 30 -6 -2026
Keywords :
Speaker recognition, Residual CNN, BiLSTM, FFT, Noisy environments.

Abstract

Despite the considerable progress made in speaker recognition technologies, maintaining robust performance under real-world noisy recording conditions remains a challenging task, as environmental noise can significantly degrade system accuracy. This work focuses on a hybrid deep learning architecture that integrates convolutional neural networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM)layers to improve speaker recognition in noisy environments. Speech signals sampled at 16 kHz are pre-processed using zero padding to ensure uniform input length, followed by feature extraction in the frequency domain. The CNN layers are used to learn distinct spectral features, and the BiLSTM layers are used to learn temporal dependencies. The proposed CNN–BiLSTM architecture was evaluated on the PCM16k dataset in noisy environments with added Gaussian white noise (AWGN) at various signal-to-noise ratios (SNR). Background noise was also added, and the CMU Arctic dataset was used to test the generalization capability of the proposed model. The experimental results show that the proposed CNN–BiLSTM model outperforms the CNN–LSTM model, achieving a high accuracy of 99.60% and maintaining robust performance even at SNR = 0 dB, where the model achieves an accuracy of 86.42%. These findings demonstrate the effectiveness of the proposed CNN–BiLSTM architecture in enhancing speaker identification performance and improving robustness against severe noise conditions.

Corresponding Author E-mail: kawther.mthem@s.uokerbala.edu.iq , ashwan.a@uokerbala.edu.iq

Peer review under responsibility of Iraqi Academic Scientific Journal and University of Kerbala.

1. Introduction

Speaker recognition is the process of analysing and processing the acoustic characteristics that are unique to each individual and can be used to verify and identify the speaker as biometric characteristics. Its importance has increased significantly with the rapid development of modern technologies, including smart devices, voice assistants, and digital authentication systems. The need for accurate and reliable verification in everyday services of speaker recognition, has become a critical element in interactions between humans and machines
and security [1].

Speaker recognition systems still face challenges in noisy environments, such as real-world environments where reverberation, channel distortion, and background noise lead to reduced system accuracy and increased error rates.[1] Pre-processing of the audio signal and feature extraction are critical for increasing the robustness of the model against noise. [1]. Deep learning techniques can be applied to enhance speaker recognition systems. Convolutional neural networks (CNNs) are used for extracting spectral features of speech signals, and memory-based models like long short-term memory (LSTM) are successfully applied for modelling temporal dependencies between speech frames. Hybrid models of convolutional neural networks (CNNs) and (LSTM) have demonstrated enhanced overall performance, especially when compared to conventional techniques like I-vectors and Gaussian mixture models (GMM) .[2] Moreover, self-lsupervised and adversarial learning techniques have been applied for enhancing the extraction of speaker embeddings from large amounts of unlabelled audio-visual data. These techniques improve the differentiability of embedding's and avoid dependence on fully labelled datasets [3]. Nonetheless, challenges remain that affect the accuracy of the model, particularly in conditions with high noise levels or when using diverse recording devices. There is a need for robust models to overcome these challenges, and these limitations lead to the design of hybrid architectures, combining CNNs and BiLSTM modules to efficiently capture spectral and temporal speech features while ensuring high accuracy and robustness in noisy environments. Figure 1 illustrates the general workflow of the speaker recognition system, showing each stage from data input to decision making. This overview of the diagram provides a clear understanding of the system.

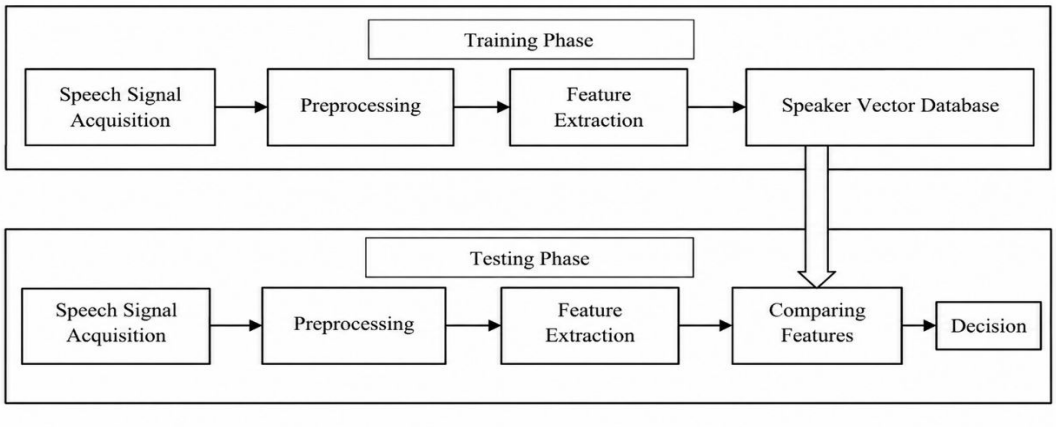

Figure: 1 Diagram of the basic steps of a speaker identification system. [4]

2. RELATED WORK

Deep learning, through the development of technology in current techniques, is gaining more popularity and attention than machine learning (ML). This is because deep learning has the ability to maintain system accuracy and handle large data sets. It is also used on real-world data. In general, machine learning requires a lot of experience and human supervision or intervention in order to implement the learning models. This provides opportunities for innovations in the learning model because it can handle a huge amount of data[5]Due to improvements in accuracy and method-logical architecture, deep learning (DL) methods have experienced a resurgence. However, current limitations of DL approaches include theoretical accumulation and unjustified DL models. Despite these shortcomings, companies that manage large da-ta sets rely heavily on DL techniques due to their superior predictive capabilities. [6]Some types of techniques are listed in Table1.

Table 1: This table shows some types of datasets and some techniques used for speak-er recognition and the results

Title	**Dataset**	**Year**	**Results**	**Techniques**	**limitations**
[7]Speaker Identification in Different Emotional States in Arabic and English	KSU Emotion & EPST dataset	2020	accuracy=97.18	CRNN, Spectrogram	The model deteriorates and becomes less accurate in emotional situations
[8] Novel Hybrid DNN Approaches for Speaker Verification in Emotional and Stressful Talking Environments	SUSAS database, and (RAVDESS)	2021	EER= 0.22 AUC = 0.99	Hybrid DNN DNN + HMM HMM +DNN DNN + GMM GMM + DNN	Needs more time to be trained
[9] CASA - Based Speaker Identification Using Cascaded GMM-CNN Classifier	Noisy +Emotional speech datasets SUSAS, ESD, RAVDESS	2021	Precision=0.82 F1scor=0.81 Recall=0.81 ROC=0.80	two different modules: a CASA Model Convolutional Neural Network Classifier for Speaker Identification.	The computational complexity of the model

[10] Augmentation vs Noise Compensation for x-vector Systems	VoxCeleb1/2	2021	improves the performance of EER. to 58%	Feature extraction (MFCC) x-vector systems	System performance decreases in the presence of noise compared to noise-free environments.
[11] Deep Learning Algorithms Based Voiceprint Recognition System in Noisy Environment	VoxForge	2021	accuracy=96%	MFCC -CNN and RW-CNN.	When dealing with raw audio, the efficiency of the system decreases.
[12] A Deep Neural Network Model for Speaker Identification	Aishell-1with Gaussian white noise added	2021	Accuracy of 91.56%	CNN, RNN→LSTM. CNN, GRU.	System accuracy deteriorates when noise is added
[13] Extended U-Net for Speaker Verification in Noisy Environments	VoxCeleb1 test set and VOiCES dataset	2022	EER =6.53	Extended U-Net, U-Net	
[14] Analysis of Speaker Identification Using YOHO Dataset	YOHO English Speech Dataset	2023	Accuracy= 91.93%	DNN/MLP, LSTM, And BLSTM. MFCC/Spectrogram	The model was only applied to clean data.

3. Dataset

In this work two datasets were used The first dataset consists of (16000_pcm_speeches) is baseline dataset previously used in related work The second dataset (CMU Arctic) is used to evaluate the generalization capability of the model.

3.1(16000_pcm_speeches) dataset

Kaggle speaker recognition dataset (16000_pcm_speeches), This dataset is divided into two folders. [15]

• The folder named background noise

contains audio files that are not speeches but can be found in and around the speaker's environment, such as audience laughter or applause. They can be mixed with the speech during training. [16]

Table 2: Summary of the Speech Dataset (16000_pcm_speeches) Used in This Study.

Attribute	Description
Source	"https://www.kaggle.com/datasets/kongaevans/speaker-recognition-dataset"
Data Type	**Speech signals** **File Format WAV**
Sampling Rate	**16,000 Hz**
Number of Speakers	**5**
Total Number of Samples	**7501**
Noise Type	**Babble noise, clapping sounds (Real-world acoustic conditions)**

• The first folder contains a dataset of samples from five speakers in interviews and speeches recorded in different outdoor environments,

3.2 (CMU Arctic) dataset

- Source: "https://www.kaggle.com/datasets/mrgabriellblins/speaker-recognition-cmu-arctic/data"
- Carnegie Mellon University (CMU)
- Type of data Speech Signals File format WAV
- This dataset consists of recordings in a clean environment and was previously used for speech analysis, but has recently been used for speaker recognition**,** uses different accents**,** it can be accessed directly**.**[17]
- The dataset contains 12,486 samples and 18 speakers

Table 3: Number of samples per speaker					
Speaker	**Samples**	**Speaker**	**Samples**	**Speaker**	**Samples**
Awe	906	haw	474	aup	474
Awb	910	axb	474	bdl	906
Clb	906	eey	475	fem	474
Gka	474	jmk	909	ksp	906
Ljm	474	lnh	905	rms	906
Rxr	533	slp	474	slt	906

4 .Proposed Methodology

4.1 System overview

The proposed speaker recognition system consists of four main steps:

pre-processing, feature extraction, deep learning of features, and classification. This pipeline works because it separates the problem into three complementary tasks,

first, presentation (FFT), make structure visible, second, Spatial learning (CNN), extract spectral cues, third, Temporal modeling (BiLSTM) capture dynamics. The following diagram explains the structure of the model.

As shown in Figure. 2, the input audio is processed and features are extracted, then passed through five residual CNN blocks to extract strong spectral features before being fed into the BiLSTM layer for temporal modelling.

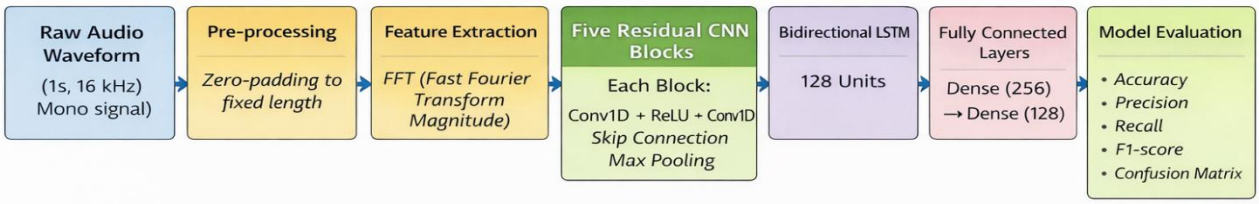

Figure 2: Architecture of the proposed 5 Residual Blocks CNN + BiLSTM speaker recognition model.

4.2 The proposed model was trained using three different training scenarios:

- In the first scenarios: the model was trained using the original audio data without adding any noise. more details are available in Section (5).
- In the second scenarios: the model was trained using audio data with added background noise to represent real-world environments. more details are available in Section (6).
- In the third scenarios: the model was trained after adding White Gaussian noise (AWGN)used different SNR 20,15,10,5,0. more details are available in Section (7).

4.3 Pre-processing and Features Extraction

Speech signals are pre-processed to extract frequency characteristics used in speaker recognition tasks. The database consists of audio recordings containing noise (Babble noise, clapping sounds), sampled at 16 kHz, and organized into folders according to speaker identity.

Initially, each audio file is loaded and converted to a single-channel (mono) signal. To ensure uniform signal length, signals longer than one second are truncated, and shorter signals are padded with zeros in the time domain to reach 16,000 samples, which is equivalent to one second of audio.

Next, the time signal is converted to the frequency domain using Fast Fourier Transform (FFT), where the magnitude spectrum is calculated and only the positive frequency components are retained due to the symmetry of the spectrum in real signals.

Finally, the extracted features for each speaker are collected and saved in compressed files in (. npz) format to facilitate the loading process, the data is divided into: training set, verification set, and test set. To prevent data leakage and ensure fair evaluation of the model,

Stratified samples are applied based on speaker identity. The random seed is set to 43 to ensure: Balanced distribution of classes

across groups. the ability to reproduce the experiment and obtain consistent results. The pre-processed data was used for speaker identification in a close set environment. Then it is entered into the model

4.4 Residual CNN Blocks

4.4.1 General Principle

Convolutional neural networks (CNNs) are effective models for extracting features from high-dimensional data such as spectral signals or sound waves, thanks to their ability to capture local patterns through convolutional filters.

However, as the depth of the network increases, it becomes difficult to maintain the gradient flow during training, which can lead to the vanishing gradient problem, where the effects of the first layers decrease during backpropagation.

- **Residual Connections**

The idea of residual networks (ResNet) was proposed to overcome this problem by using residual connections that directly connect the layer's input to its output. The basic meth-od is to allow the block to learn the residual difference between the input and output in-stead of learning the complete transformation.

Its basic formula is:

(x)= H(x)-x You learn only the differences.

H(x)=$F(x)+x$

Where: x, x is the input to the block(x)

F(x) is the representation learned by the layers within the block

H(x) Represents the final output

Thanks to this structure, deeper networks can be trained with higher efficiency, improving the extraction of subtle acoustic features relevant to the speaker. [18]

Function of layers within the block:

• 1D Convolution: This extracts local spatial or temporal patterns from the spectral signal.

• Activation Functions: These include ReLU, which introduces non-linearity to the model, allowing it to learn non-linear relationships.

• Max Pooling: Max Pooling: This reduces dimensions, allowing the model to concentrate on the most significant features while also reducing complexity. This combination of layers enables the model to learn hierarchical features that are valid for more stable speaker identification.

4.5 BiLSTM–Bidirectional Long Short-Term Memory Layer

Recurrent Neural Networks (RNNs) are specifically designed for handling sequential data, where the states retain a partial memory of the former timeline, which is a significant aspect of signal sequences, like audio.

However, the basic RNN model has difficulty in understanding long-term temporal

dependencies because of the vanishing gradient problem. Hence, the LSTM (Long Short-Term Memory) network was proposed, which incorporated gates (input, forget, output gates) to delay the information flow.

4.5.1 Bidirectional processing

Instead of processing sequences in the traditional front-to-back manner, BiLSTM uses two networks:

- The first moves backward from the past to the future
- The second moves forward from the future to the past

This technique provides rich contextual information in both directions, increasing the model's ability to recognize long-term acoustic patterns.

4.6Classification Layer

The output of the BiLSTM layer is fed into fully connected dense layers, followed by a softmax output layer. The softmax function produces posterior probabilities for each speaker class, enabling multi-class speaker identification.

4.7Model Optimization

Adam and early stopping were used.
Early stopping prevents overfitting and ensures generalization, while Adam reduces the loss function by adjusting learning rates.

4.8 Training Configuration and Parameters

In this section, all parameters used in model construction are explained. Filters are mentioned, as well as the number of blocks, etc. To evaluate the robustness of the proposed system under noisy conditions, additive white Gaussian noise (AWGN) is introduced at SNR levels. The performance of the system is evaluated using multiple evaluation metrics.

Table 4. Training Configuration and Parameters

Category	Parameter	Value
Data Split	Training / Validation / Testing	80% / 10% / 10%, seed=43
Loss	Loss function	Sparse -Categorical- cross entropy
CNN Architecture	Number of residual blocks	5
CNN Filters	Filters per block	16, 32, 64, 128, 128
Kernel Size	Convolution kernel	3
Activation Function	CNN & Dense layers	ReLU
Pooling	MaxPooling size	2
LSTM Type	Type & Units	Bidirectional, 128 units
Dense Layers	Fully connected layers	256, 128
Output Layer	Activation function	Softmax
Optimizer	Optimization algorithm	Adam
Batch Size	Samples per batch	128
Epochs	Maximum epochs	100
Early Stopping	Patience	10 epochs
Noise	AWGN SNR levels	0, 5, 10, 15, 20 dB

5. Results: PCM dataset

The proposed model was tested and trained using the dataset (of 16,000 pcm). The model contained audio recordings taken in a noisy environment and had a test accuracy of 99.60% and a test loss of 0.0171. The error rate had a value of 0.0053. The AUC values for all speaker classes ranged from 0.9999 to

1.0000, with overall AUC scores of 1.0000. This confirms the robustness of the model and its applicability in real-world environments.

5.1. :Accuracy curve

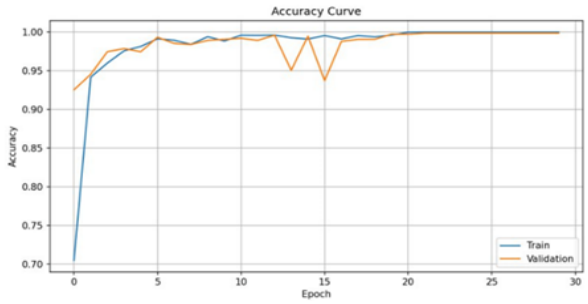

Figure 3: Accuracy curve of the CNN–BiLSTM model on the original PCM dataset

The graph shows how the accuracy of the model improved over time, The temporary decline observed at epoch 13 can be attributed to a transient fluctuation in the Adam optimization algorithm while balancing the weights between the spatial features extracted by the CNN and the temporal context modeled by the BiLSTM. The model quickly overcame this fluctuation through self-correction, with the model reaching peak accuracy after approximately 15–20 training epochs.

The training and validation curves also indicate that the model does not over fit.

5.2. The loss curve

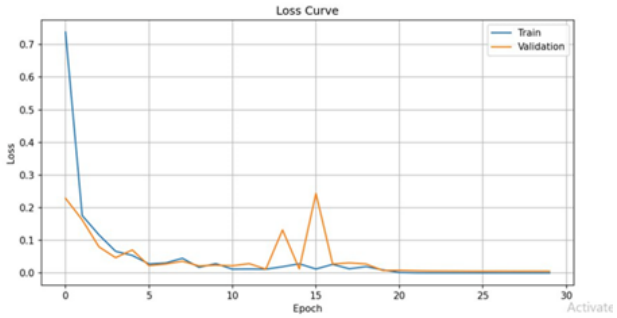

Figure 4: loss curve of the CNN–BiLSTM model on the original PCM dataset.

The loss curve shows how the loss drops over time (epochs) for both training and validation data. The model was trained for approximately 30 epochs and reached a stable point after about 20 epochs.

5.3. Receiver operating characteristic

Creating a ROC curve contributes to a deeper evaluation of the classification model's performance, as it relies on calculating both the true positive rate (TPR) and the false

positive rate (FPR) at different decision threshold levels. The values are obtained using the ROC_curve function of the metrics module in the scikit-learn library. The AUC coefficient, extracted via the AUC function of the same module, is used to measure the area under the ROC curve. The true positive rate is plotted on the vertical axis, and the false positive rate on the horizontal axis, with the curve representing the relationship between them across different classification thresholds.

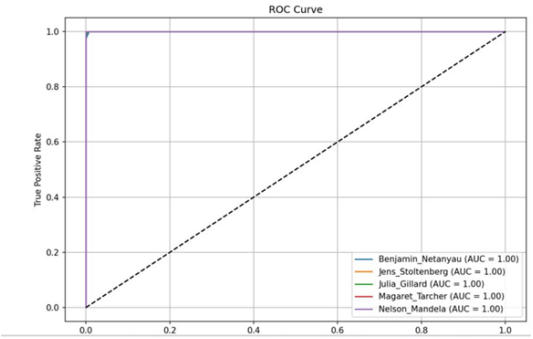

Figure. 5: ROC curve of the CNN–BiLSTM model on the original PCM dataset.

The proposed model achieved an AUC of 1.00, ✓ indicating excellent discrimination capability.

The black dotted line shows random performance (AUC = 0.5), meaning that any point on this line indicates that the correctly predicted proportion (true positive rate) is exactly equal to the incorrectly predicted proportion (false positive rate).

AUC per speaker:

Table 5: Area Under the Curve (AUC) Values for Each Speaker and Overall Model Performance

Speaker	AUC	Speaker	AUC
Benjamin_Netanyau	0.9999	Jens_Stoltenberg	1.0000
Julia_Gillard	1.0000	Magaret_Tarcher	1.0000
Nelson_Mandela	1.0000	**Total AUC**	**1.0000**

5.4. equal error rate (EER):

EER is one of the most commonly used metrics in speaker recognition tasks, and EER is the point at which (FAR = FRR). EER is determined using the mathematical equation (1): [19]

$$EER = \frac{1}{2}\left(\frac{FAR+FRR}{2}\right) \quad (1)$$

Where in Equation (1), By calculating EER, we will be able to identify the places where the system is likely to commit both types of errors (FRR and FAR) to the same degree. [19]

Error Rate= 0.004: This rate is considered very low and is excellent in most classifications, indicating that the model has a high predictive power.

5.5. confusion matrix

The confusion matrix represents the performance of the speaker classification model. This graph compares the actual/true categories with the predicted categories for five political figures.

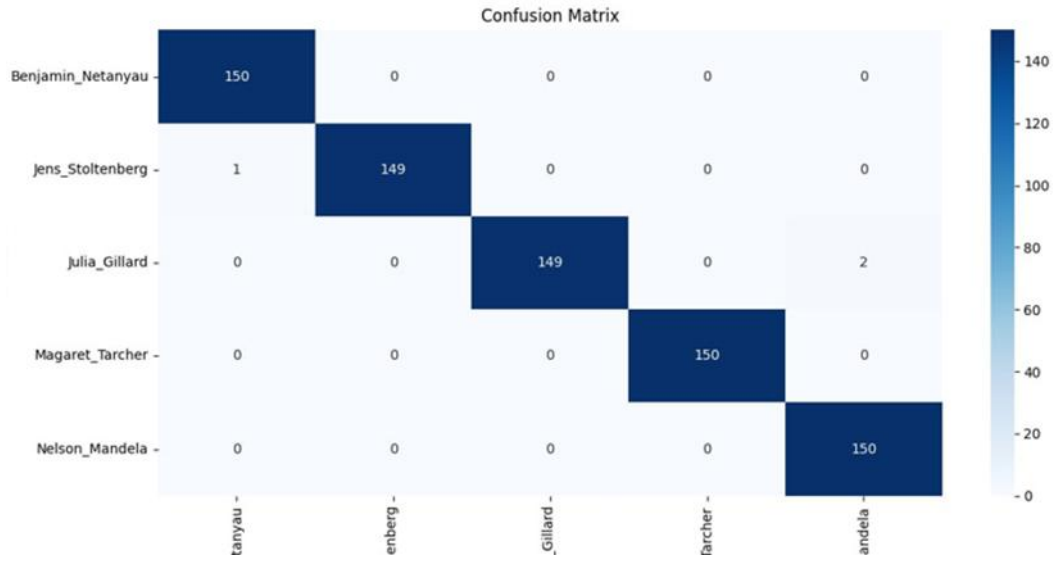

Figure 6: confusion matrix of the CNN–BiLSTM model on the original PCM dataset.

Overall performance The concentration of high numbers on the main axis indicates that the model has high accuracy and good ability to distinguish between different speakers.
Model performance analysis:
Correct predictions: Large numbers on the main diagonal (dark blue cells) indicate true positives. For example, 150 instances of Nelson Mandela's voice were correctly classified as Nelson Mandela.
False predictions: Numbers outside the main cluster are misclassifications (False Positives/Negatives). The model shows very few errors, such as incorrectly classifying 3 instances of Benjamin Netanyahu's voice as Jens Stoltenberg.
..

6. Model Evaluation after add background noise

Noise was added to the background using noise clips extracted from the original dataset, and these clips were segmented into one-second intervals to match the length of the audio samples, in order to evaluate the robustness of the model

Table 6: To illustrate the performance evaluation of the speaker recognition model using two types, CNN–LSTM and CNN–BiLSTM, on the original PCM data and data with added background noise.

Dataset / Noise	Model	Test Accuracy (%)	Test Loss	Precision (VA)	Recall (VA)	F1-score (VA)	Incorrect Predictions	AUC	EER (%)
Original (PCM)	"CNN–BiLSTM "	99.60	0.014	0.9960	0.9960	0.9960	3	0.9998	0.65
Original (PCM)	" CNN–LSTM"	98.14	0.0453	0.9815	0.9814	0.9813	14	0.981	1.86%
Background Noise	"CNN–BiLSTM "	95.07	0.198	0.9510	0.9507	0.9506	37	0.9860	4.20
Background Noise	" CNN–LSTM"	89.35	0.4830	0.8938	0.8934	0.8926	80	0.988	5.66

From the table, it can be seen that the CNN–BiLSTM model consistently performs better than the CNN+ LSTM model, both on the original PCM16k dataset and on data with added background noise. The CNN–BiLSTM model achieves higher evaluation metrics, demonstrating greater robustness and generalization ability even under noisy conditions The results indicate that the CNN–BiLSTM model achieved the best performance in the case of the original data (PCM1), with an accuracy of 99.60%, as well as high Precision, Recall, and F1 scores, all of which reached 0.9960, and a significant decrease in the number of false predictions (only 3). This reflects the model's ability to represent temporal characteristics in both forward and backward directions with high efficiency.

Table7: The table shows the effect of Adam and skip connection on performance.

Dataset / Noise	Model	Test Accuracy (%)	Precision (VA)	Recall (VA)	F1-score (VA)	Incorrect Predictions
Background Noise	"CNN–LSTM " used module without skip connection	87.75	0.8764	0.8774	0.8761	92
Background Noise	"CNN–LSTM " used module without skip connection + Adam	72.97	0.7391	0.7297	0.7292	203

As for the CNN–LSTM model without Skip Connection, it recorded the lowest performance (87.75%), with an increase in the number of false predictions to 92, indicating that removing the Skip Connection mechanism negatively affected

the flow of information within the network and weakened its ability to retain important features, especially in the presence of noise. This indicates that the use of CNN residual block is stronger than CNN block.
As for the CNN–LSTM model without Skip Connection and Adam lower perform after add Adam accuracy high 72.97 to 87.75
of the model compared to other architectures
.

Overall, the results show that integrating CNN with BiLSTM with skip connection with Adam provides superior performance in speaker recognition tasks, both in clean and noisy conditions, and enhances the robustness and noise resistance

Table 8: Comparison of the proposed hybrid model and previous work on the same data set

Reference	Year	Dataset	Model	Type of noise	Results
[20]An Ensemble Approach for Speaker Identification from Audio Files in Noisy Environments	2024	PCM 16000	BiLSTM	Only noise in dataset	Test accuracy=0.929 Test Loss=0.263 Test Mean squared error=0.0218
[16]An investigation into the reliability of speaker recognition schemes: analysing the impact of environmental factors utilising deep learning techniques	2024	PCM 16000	CNN model	add background noise	ROC(areas=0.83)
[21]Deep Learning and Fourier Transform for Speaker Recognition(DLFSR)	2025	PCM 16000	CNN model	Only noise in dataset	accuracy=0.9871
Proposed methodology	2026	PCM 16000	CNN–BiLSTM	Only noise in dataset , Add background noise	accuracy= 0.9960 loss= 0.0171. MSE= 0.001916 ROC (areas=1.0000) Error Rate= 0.004 Add background noise accuracy= 0.9507 ROC (areas=0.9860)

The table provides a comparative analysis of previous studies and the proposed CNN-BiLSTM model for speaker recognition, focusing on the dataset used, model type, noise type, and key performance metrics. The com

parison demonstrates the superior performance and high robustness of the proposed approach in both the original data environments and after adding background noise.

7. Model Evaluation under Different SNR Levels

White Gaussian noise (AWGN) was added to the original dataset, and we now have new data for each level at different signal-to-noise ratio (SNR) levels of 0, 5, 10, 15, and 20 dB The noisy data was then divided and entered into the model in order to test the model's robustness against noise.

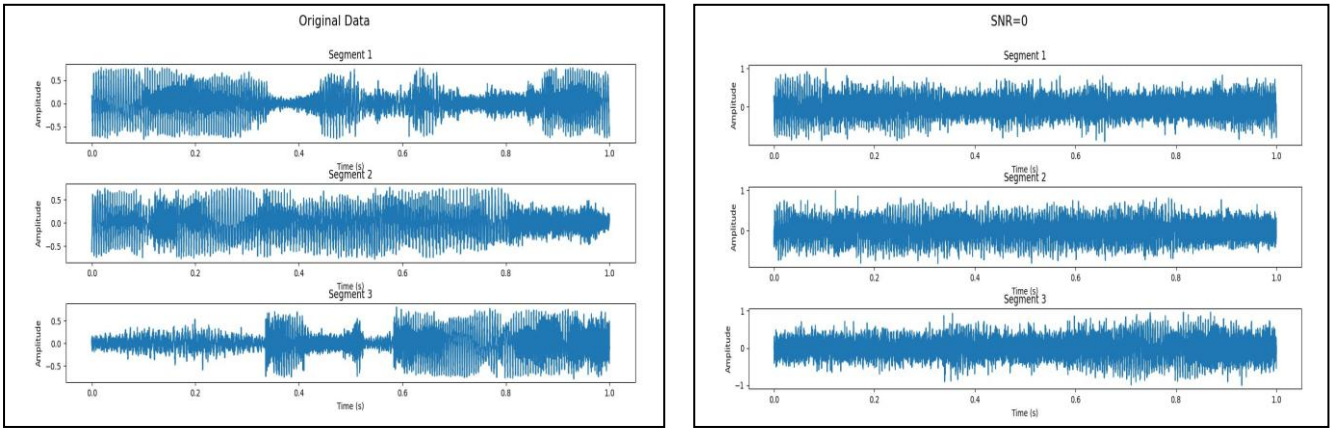

Figure 7: Through this graph, we can see how the audio signal was distorted after adding snr=0.

7.1. Classification Report:

P: precision
R: recall
F1: f1-score

Table 9: Report of classification by speakers (P, R, F1) across different SNR levels

Speaker	Support	Original Noisy Speech	SNR 20	SNR 15	SNR 10	SNR 5	SNR 0
"Benjamin Netanyahu"	150	P:0.9934 R:1.0000 F1:0.9967	P:0.9932 R:0.9733 F1:0.9832	P:0.9655 R:0.9333 F1:0.9492	P:0.9645 R:0.9067 F1:0.9347	P:0.8987 R:0.9467 F1:0.9221	P:0.8231 R:0.8067 F1:0.8148
"Jens Stol-tenberg"	150	P:1.0000 R:0.9933 F1:0.9967	P:0.9671 R:0.9800 F1:0.9735	P:0.9600 R:0.9600 F1:0.9600	P:0.9664 R:0.9600 F1:0.9632	P:0.9589 R:0.9333 F1:0.9459	P:0.8269 R:0.8600 F1:0.8431
"Julia Gillard"	151	P:1.0000 R:0.9868 F1:0.9933	P:0.9934 R:1.0000 F1:0.9967	P:0.9934 R:1.0000 F1:0.9967	P:0.9867 R:0.9801 F1:0.9834	P:0.9600 R:0.9536 F1:0.9568	P:0.8024 R:0.8874 F1:0.8428
"Margaret Thatcher"	150	P:1.0000 R:1.0000 F1:1.0000	P:0.9868 R:0.9933 F1:0.9900	P:0.9613 R:0.9933 F1:0.9770	P:0.9255 R:0.9933 F1:0.9582	P:0.9667 R:0.9667 F1:0.9667	P:0.9034 R:0.8733 F1:0.8881
"Nelson Mandela"	150	P:0.9868 R:1.0000 F1:0.9934	P:1.0000 R:0.9933 F1:0.9967	P:1.0000 R:0.9933 F1:0.9967	P:0.9867 R:0.9867 F1:0.9967	P:0.9932 R:0.9733 F1:0.9832	P:0.9853 R:0.8933 F1:0.9371

Table 10: Performance of the speaker recognition model in noisy conditions with different levels of SNR (PCM dataset)

Condition	SNR (dB)	Test Accu-racy (%)	Precision (VA)	Recall (VA)	F1-score (VA)	Correct Pre-dictions	Incorrect Pre-dictions
Original Noisy Speech	……	99.60	0.9960	0.9960	0.9960	748	3
Noisy	20	98.67	0.9867	0.9867	0.9867	741	10
Noisy	15	97.60	0.9760	0.9760	0.9759	733	18
Noisy	10	96.54	0.9660	0.9654	0.9652	725	26
Noisy	5	95.47	0.9555	0.9547	0.9549	717	34
Noisy	0	86.42	0.8682	0.8642	0.8652	649	102

Table 11: compares the performance of the CNN–BiLSTM on PCM 16000 and CMU Arctic data under different SNR) conditions and original data.

Table 11: SI performance across multiple datasets under different SNR

Condition / SNR	Dataset	Accuracy (%)	Precision (Macro)	Recall (Macro)	F1-score (Macro)	Correct Pre-dictions	Incorrect Predictions
Original	PCM 16000	99.60	0.9960	0.9960	0.9960	748	3
Original	CMU Arctic	97.52	0.9751	0.9752	0.9749	1218	31

SNR 20	PCM 16000	98.67	0.9867	0.9867	0.9867	741	10
SNR 20	CMU Arctic	94.00	0.9408	0.9400	0.9395	1174	75
SNR 15	PCM 16000	97.60	0.9760	0.9760	0.9759	733	18
SNR 15	CMU Arctic	93.43	0.9246	0.9180	0.9203	1167	82
SNR 10	PCM 16000	96.54	0.9660	0.9654	0.9652	725	26
SNR 10	CMU Arctic	89.75	0.8866	0.8745	0.8777	1121	128
SNR 5	PCM 16000	95.47	0.9555	0.9547	0.9549	717	34
SNR 5	CMU Arctic	87.51	0.8610	0.8491	0.8475	1093	156
SNR 0	PCM 16000	86.42	0.8682	0.8642	0.8652	649	102
SNR 0	CMU Arctic	83.51	0.8142	0.8160	0.8109	1034	206

The results show that the accuracy of the model is reduced when noise is increased through different SNRs, which is expected.
However, it gave strong results and did not collapse even at SNR=0, which indicates that it is a model that can be used in the real world.

8. Discussion

The experimental results showed that the proposed model based on the integration of CNN + BiLSTM achieved better accuracy and noise resistance compared to the traditional CNN + LSTM model. Replacing the unidirectional LSTM layer with a bidirectional BiLSTM layer resulted in a significant improvement in classification accuracy when testing the model on the original data. This improvement is due to BiLSTM's ability to extract temporal dependencies from both directions (forward and backward), allowing for a more accurate representation of the acoustic signal characteristics.
The results also showed that adding Skip Connection (Residual Blocks) within the CNN layers greatly contributed to improving feature extraction and reducing information loss during deep passing. The model based on the residual architecture (Residual CNN) outperformed the traditional CNN model, confirming that the use of skip connections helps stabilize the training process and enhances the model's generalization ability.
Since there was no database available with all types of real-world noise, we introduced some background noise to test the robustness of the model in adverse environments. Though the original dataset had environment noise, additive white Gaussian noise (AWGN) was introduced at varying levels of signal-to-noise ratio (SNR = 20, 15, 10, 5, 0 dB) to test the model's robustness in adverse environments.
As expected, the performance gradually reduced with the increase in noise levels, but the reduction was marginal and did not go to the extent of failure even at SNR = 0 dB The minimum accuracy achieved was 86.42%, which is a clear indication of the robustness of the model compared to the previous studies, where the performance drastically reduced at low noise levels.
On the other hand, the second dataset, which was slightly imbalanced, also showed slightly reduced performance, but the values were within acceptable limits, indicating the model's capability to handle class imbalance without much loss of accuracy.
The results clearly indicate that the integration of the residual structure of CNNs and BiLSTM layers offers a robust framework for speaker recognition in noisy audio environments.

9. Conclusion

In this study, a robust speaker identification system was developed by integrating Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) architectures to effectively extract both spectral and temporal characteristics from speech signals. The CNN layers were employed to learn discriminative spectral features from FFT-based representations, while the BiLSTM layers captured long-term temporal dependencies within speech sequences, resulting in improved speaker discrimination capability.

To evaluate the effectiveness of the proposed approach, extensive experiments were conducted on the original PCM speech dataset and under various noisy conditions generated by adding artificial noise at different Signal-to-Noise Ratio (SNR) levels. Furthermore, the CMU Arctic dataset was utilized to assess the generalization capability of the model on unseen speech data. Experimental results demonstrated that the proposed CNN–BiLSTM model consistently outperformed the baseline CNN–LSTM architecture across different testing scenarios. The model achieved an accuracy of 99.60% on the original dataset and maintained strong robustness under severe noise conditions, achieving an accuracy of 86.42% even at an SNR level of 0 dB.

These findings confirm that the integration of CNN and BiLSTM layers provides a powerful framework for speaker identification in noisy environments by effectively learning complementary spectral and temporal information. Moreover, the results indicate that the proposed model can maintain reliable performance despite significant degradation in speech quality caused by environmental noise.

Future work may focus on evaluating the proposed architecture using larger and more diverse speech corpora, incorporating real-world environmental noise sources, and investigating advanced feature extraction and data augmentation techniques to further improve system robustness and generalization performance.

10.Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, Aug. 2021, doi: 10.1016/j.neunet.2021.03.004.

[2] S. K. Nayak, A. K. Nayak, S. R. Laha, N. Tripathy, and T. A. Smadi, "A Robust Deep Learning-Based Speaker Identification System Using Hybrid Model on KUI Dataset," *Int. J. Electr. Electron. Res.*, vol. 12, no. 4, pp. 1502–1507, Dec. 2024, doi: 10.37391/ijeer.120446.

[3] H. Tao, Ruijie and Lee, Kong Aik and Das, Rohan and Hautamäki, Ville and Li, "Self-supervised Speaker Recognition with Loss-gated Learning Title," 2021, doi: 10.48550/arXiv.2110.03869.

[4] T. S. Mohammed, K. M. Aljebory, M. A. Abdul Rasheed, M. S. Al-Ani, and A. M. Sagheer, "Analysis of Methods and Techniques Used for Speaker Identification, Recognition, and Verification: A Study on Quarter-Century Research Outcomes," *Iraqi J. Sci.*, no. 0067–2904, pp. 3256–3281, Sep. 2021, doi: 10.24996/ijs.2021.62.9.38.

[5] N. Shome, A. Sarkar, A. K. Ghosh, R. H. Laskar, and R. Kashyap, "Speaker Recognition through Deep Learning Techniques," *Period. Polytech. Electr. Eng. Comput. Sci.*, vol. 67, no. 3, pp. 300–336, Jul. 2023, doi: 10.3311/PPee.20971.

[6] Q. Le, L. Miralles-Pechuán, S. Kulkarni, J. Su, and O. Boydell, "An Overview of Deep Learning in Industry," in *Data Analytics and AI*, Auerbach Publications, 2020, pp. 65–98. doi: 10.1201/9781003019855-5.

[7] A. H. Meftah, H. Mathkour, S. Kerrache, and Y. A. Alotaibi, "Speaker Identification in Different Emotional States in Arabic and English," *IEEE Access*, vol. 8, pp. 60070–60083, 2020, doi: 10.1109/ACCESS.2020.2983029.

[8] I. Shahin, A. B. Nassif, N. Nemmour, A. Elnagar, A. Alhudhaif, and K. Polat, "Novel hybrid DNN approaches for speaker verification in emotional and stressful talking environments," *Neural Comput. Appl.*, vol. 33, no. 23, pp. 16033–16055, Dec. 2021, doi: 10.1007/s00521-021-06226-w.

[9] 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," *Appl. Soft Comput.*, vol. 103, p. 107141, May 2021, doi: 10.1016/j.asoc.2021.107141.

[10] M. Mohammad Amini and D. Matrouf, "Data augmentation versus noise compensation for x-vector speaker recognition systems in noisy environments," in *2020 28th European Signal Processing Conference (EUSIPCO)*, IEEE, Jan. 2021, pp. 1–5. doi: 10.23919/Eusipco47968.2020.9287690.

[11] H. Y. Khdier, W. M. Jasim, and S. A. Aliesawi, "Deep Learning Algorithms based Voiceprint Recognition System in Noisy Environment," *J. Phys. Conf. Ser.*, vol. 1804, no. 1, p. 012042, Feb. 2021, doi: 10.1088/1742-6596/1804/1/012042.

[12] F. Ye and J. Yang, "A Deep Neural Network Model for Speaker Identification," *Appl. Sci.*, vol. 11, no. 8, p. 3603, Apr. 2021, doi: 10.3390/app11083603.

[13] J. Kim, J. Heo, H. Shim, and H.-J. Yu, "Extended U-Net for Speaker Verification in Noisy Environments," Jun. 2022, doi: https://doi.org/10.48550/arXiv.2206.13044.

[14] N. M. Almarshady, A. A. Alashban, and Y. A. Alotaibi, "Analysis and Investigation of Speaker Identification Problems Using Deep Learning Networks and the YOHO English Speech Dataset," *Appl. Sci.*, vol. 13, no. 17, p. 9567, Aug. 2023, doi: 10.3390/app13179567.

[15] P. Budiga, B. B, G. Gunisetty, N. D. Moka, and G. V. S. Reddy, "CNN Trained Speaker Recognition System in Electric Vehicles," in *2022 International Virtual Conference on Power Engineering Computing and Control: Developments in Electric Vehicles and Energy Sector for*

Sustainable Future (PECCON), IEEE, May 2022, pp. 1–6. doi: 10.1109/PECCON55017.2022.9851029.

[16] O. R. Khazaleh and L. A. Khrais, "An investigation into the reliability of speaker recognition schemes: analysing the impact of environmental factors utilising deep learning techniques," *J. Eng. Appl. Sci.*, vol. 71, no. 1, p. 13, Dec. 2024, doi: 10.1186/s44147-023-00351-0.

[17] A. W. Black, "CMU Wilderness Multilingual Speech Dataset," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2019, pp. 5971–5975. doi: 10.1109/ICASSP.2019.8683536.

[18] K. H. And, X. Z. And, S. R. And, and J. Sun, "Deep Residual Learning for Image Recognition," *CoRR*, vol. abs/1512.0, 2015, doi: 10.48550/arXiv.1512.03385.

[19] J. C. Cavalcanti, R. R. da Silva, A. Eriksson, and P. A. Barbosa, "Exploring the performance of automatic speaker recognition using twin speech and deep learning-based artificial neural networks," *Front. Artif. Intell.*, vol. 7, Feb. 2024, doi: 10.3389/frai.2024.1287877.

[20] S. S. Zarin, E. Mustafa, S. K. uz Zaman, A. Namoun, and M. H. Alanazi, "An Ensemble Approach for Speaker Identification from Audio Files in Noisy Environments," *Appl. Sci.*, vol. 14, no. 22, p. 10426, Nov. 2024, doi: 10.3390/app142210426.

[21] T. M. Sayed, A. Gody, and S. T. Muhammad, "Deep Learning and Fourier Transform for Speaker Recognition(DLFSR)," *Fayoum Univ. J. Eng.*, vol. 8, no. 1, pp. 143–151, Jan. 2025, [Online]. Available: https://fuje.journals.ekb.eg/article_411083.html