Convolutional Neural Network with Long Short-Term Memory for Speaker Identification and Verification System

Sukumar B S[1], Dr. G N Kodanda Ramaiah [2], Dr. Sarika Raga [3], Dr.Lalitha Y S [4], Dr. G K Venkatesh[5]

[1]Research scholar, Department of ECE, Visvesvaraya Technological University, Belagavi, Karnataka, India
[2]Professor and HOD Department of ECE, Kuppam College of Engineering, Kuppam, Andhra Pradesh, India
[3]Associated Professor, Department of ECE, Visvesvaraya Technological University PG Centre, Muddenahalli, Chikkaballapura, India
[4]Professor Department of Electronics and Communication, Don Bosco Institute of Technology, Bengaluru, Karnataka, India
[5] Professor and HOD Department of ECE, C.Byregowda Institute of Technology, Kolar, Karnataka, India

Abstract

The speaker recognition is an advanced technique to identify individuals based on unique biometric characteristics in their speech patterns. The challenge is to enhancing noise robustness and performance of speaker verification systems. To tackle this problem, this research proposed Convolutional Neural Network with Long Short-Term Memory (CNN-LSTM) for speaker identification and verification system. The CNN extract local patterns and LSTM is better at handling sequential data and captures long-term dependencies. It enhances effective and reliable speaker verification system in real-world applications which contributes accurate and secure biometric authentication system. The collected dataset is preprocessed by label encoding which converts categorical speaker identities into numerical values. Then, Mel-Frequency Cepstral Coefficient (MFCC) is used for feature extraction which efficiently represent speech signal's power spectrum and widely used for speaker recognition. The MFCC captures significant characteristics of human voice which are crucial for distinguishing among speakers. The proficiency of CNN-LSTM is assessed with evaluation metrics of accuracy, f1-score, recall and precision. The CNN-LSTM attained high accuracy 99.09%, f1-score 99.09%, recall 99.09%, and precision 99.09% which shows better performance than existing classifiers such as CNN and Bidirectional LSTM (BiLSTM).

Keywords: Convolutional Neural Network, Label Encoding, Long Short-Term Memory, Mel-Frequency Cepstral Coefficient, Speaker Recognition and Verification.

1. Introduction

Nowadays, the speaker verification system is automated for enhancing authenticity score of digital applications which has been one of the significant types of research over five decades [1]. In speaker verification, the voice of verified speaker is associated with claimed person voice and system finished with the decision to reject or accept process through threshold [2]. In speaker identification, the system tried to recognize unknown person voice among group of persons [3]. It has two various ways such as text-dependent and text-independent [4-6]. The text-dependent system is fully concentrated with voice characteristics whereas independent system not only concentrated on voice attributes [7]. The speaker identification has three primary stages such as feature extraction, modeling and scoring [8-10]. The speech signal is changing continuously because of speech activities and signal is divided into small frames where signal is constant and each frame indicated by spectral feature vectors [11]. The speaker model is trained through speaker-specific features and stored in database [12]. At verification stage, utterance is scored over stored claimed speaker and final decision is calculated to reject or accept system access [13].

The voice signal includes numerous instances and data which is used to extract data about expression, speech words, styles, emotion, accent, speaker identity, gender, health state, age of speaker and so on [14, 15]. The integration of gender data in the growth and testing process provides trustworthy information. The neural network attains another element for recognizing task-specific voice quality of two genders [16-18]. The gender verification through audio signal is essential for numerous applications such as advertising voice assistants, targeted answers, population statistics through age analysis using speech data helps in illegal examination [19]. The problem is to develop robust speaker verification system that accurately authenticates speaker identity through voice even in noisy environment with mismatched training and testing conditions. The challenge is to enhancing noise robustness and performance of speaker verification systems. This research proposed CNN-LSTM for speaker identification and verification system. The CNN extract local patterns in data whereas LSTM is better at handling sequential data and captures long-term dependencies. It enhances reliable and effective speaker verification system in real-world applications which contributes accurate and secure authentication system. The contributions of the research are described below:

- The MFCC which efficiently represent speech signal’s power spectrum and widely used for speaker recognition. It captures significant characteristics of human voice which are crucial for distinguishing among speakers.
- The CNN extract local patterns and LSTM is better at handling sequential data and captures long-term dependencies. It improves reliable and efficient speaker verification system in real-world applications.

The remaining portion of the research is organized as follows: Section 2 analyzes literature review; Section 3 provides details of proposed methodology; Section 4 gives results and discussion and section 5 concludes the research.

2. Literature Review

Several Machine Learning (ML) and Deep Learning (DL)-based classifiers have been used for speaker verification. The recent research in speaker verification system are analyzed in this section.

Ugur Ayvaz *et al.* [20] developed an automatic speaker recognition through MFCC with Multi-Layer Perceptron (MLP). The MFCC is a sequence of voice signal-specific features and human perception was non-linear after filter bank in MFCC and transferred to obtain log filter banks feature-based spectrograms through using Discrete Cosine Transform (DCT). At MFCC process, high correlated features are extracted and given to ML algorithms. The MLP provides non-linearity by activation functions which enables to capture difficult patterns within the data. However, it prone to overfitting when model was complex and numerous parameters which leads poor performance in speaker verification system.

J. V. Thomas Abraham *et al.* [21] suggested a DL for speaker identification through MFCC and CNN. The MFCC features are augmented through Chroma Energy Normalized Statistics (CENS) features to train CNN model. The cepstral through chroma features are extracted from input signal and CNN was trained with extracted features. The CNN was trained to discriminate among speakers and unknown speech signals are predicted from group of known speakers. The CNN effectively extract local features because of its convolutional layers that assist to recognize speaker characteristics from speech data. However, It unable to capture long-term temporal dependencies in speech data which decreases the model performance.

Tsung-Han Tsai and Tran Dang Khoa [22] implemented a Dual-Sequences Gate Attention Unit (DS-GAU) for speaker verification. Dual inputs from same source are considered as pooling layer in x-vector and frame layer data in x-vector. It was implemented through applying attention mechanism into Gated Recurrent Unit (GRU) for enhancing model capability. It transferred by DS-GAU to integrate much data from different temporal context of input feature while frame training. The DS-GAU process both global and local features and capturing fine-grained information which leads accurate speaker verification. However, it difficult to train and reduces the performances due to its inability to handle temporal dependencies.

Abeer Ali Alnuaim *et al.* [23] presented a Deep Neural Network and ResNet50 (DNN-ResNet50) for speaker recognition. The DNN-ResNet50 was trained to select necessary data from speech spectrograms for classification layer which perform recognition. It determined if spectrograms provide adequate information for accurate speaker classification. The ResNet50 was choose required data from speech spectrograms for classification layer for performing speaker recognition. The DNN-ResNet50 was easy to manage high-dimensional data that creates suitable for solving complex problem. However, it reduces performance and noise robustness with mismatched training and testing conditions.

Kodali Radha and Mohan Bansal [24] introduced a Bidirectional LSTM (BiLSTM) for closed-set speaker identification. The multi-scale wavelet scattering transform was utilized for higher-frequency loss data which was caused through MFCC technique. The large-scale speaker identification through using wavelet scattered BiLSTM. After training BiLSTM on wavelet feature coefficient extracted from audio files and used to predict speakers in individual test set. The BiLSTM captured long-term dependencies in speech data. However, it leads overfitting because the model learns irrelevant and noisy pattern in training data which resultant in poor performance.

From the above analysis the existing methods have drawbacks like prone to overfitting due to its numerous and complex parameters, learns noisy and irrelevant patterns in training data. It was unable to capture long-term temporal dependencies, complex to train due to its inability to handle temporal dependencies which provides poor performance. It reduces performance and noise robustness with mismatched conditions of training and testing. According to this analysis, this research proposed an CNN-LSTM for speaker identification and verification system.

3. Proposed Methodology

The CNN-LSTM is proposed in this research for speaker verification from the collected dataset. The dataset is preprocessed by label encoding which converts categorical speaker identities into numerical values. Then, the MFCC is used to captures significant characteristics of human voice which are crucial for distinguishing among speakers. The extracted features are provided to CNN-LSTM for identification and verification due to its capability to capture both spatial and temporal features of speech signal. Figure 1 shows the process of proposed methodology.

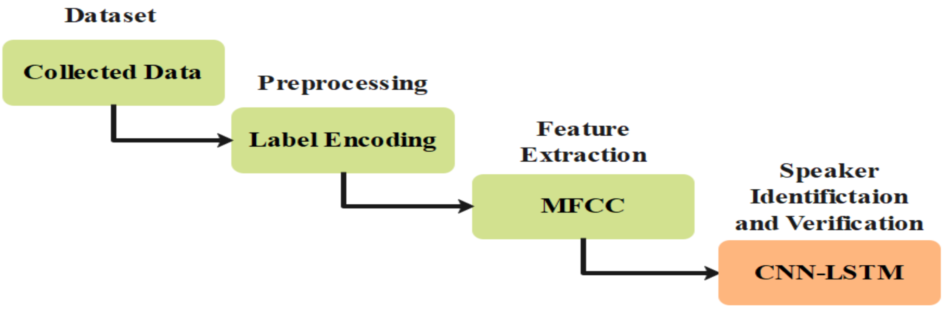

Figure 1. Process of proposed methodology

3.1 Dataset

The dataset is collected from PRAAT software which contains 13 speakers/classes with 50 samples. This dataset contains 9 males and 4 female samples with frequency range of 16khz and it is divided based on training and testing sets in the ratio of 80:20. Totally, this dataset contains 31790 files and 468 folders.

3.2 Preprocessing

The collected dataset is preprocessed by label encoding which converts categorical speaker identities into numerical values. Unlike one-hot encoding, which enhances the data dimensionality with huge number of features whereas label encoding allocates single integer value to every category. The label encoding uses nominal variables in which all the variables contain infinite set of discreate classes with no connection among classes.

3.3 Feature Extraction

The preprocessed data features are extracted by using MFCC which efficiently represent speech signal’s power spectrum and widely used for speaker recognition. It captures significant characteristics of human voice which are crucial for distinguishing among speakers. Figure 2 denotes the process of MFCC based feature extraction.

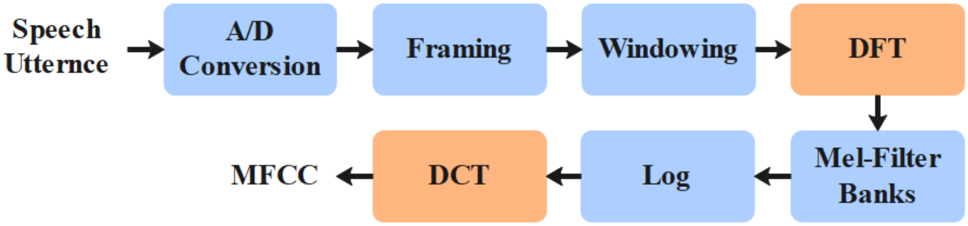

Figure 2. MFCC based feature extraction

A/D Conversion: The normal speech signal is analog signal that is changed into digital through minimum sampling rate of 16kHz. Additionally, it changes continuous speech wave into discrete signals [25].

Framing: The audio signal is continuously converted which is essential to split input signal to segments for consistent audio features known as frames. Generally, the frame size is 20-30ms ranges, as its value is greater or lesser affects the dependability of spectral calculation. The main aim of frame overlapping is to handle signal smoothness and feature invariance.

Windowing: The windows is multiplied through each frame for maintaining continuousness signal. It minimizes the spectral distortion and discontinuity of speech signal for an extent.

DFT: It is primary step for spectral analysis of feature extraction which converts time domain signal into frequency domain as eq. (1),

$$X(i) = \sum_{k=0}^{N-1} x(k) e^{\frac{-j2\pi ki}{N}} \tag{1}$$

Where, N is number of points in time domain to frequency domain and $0 \leq k \leq N-1$.

Mel-Filter Banks: The DFT signal is enabled to transfer by Mel-filter bank for obtain Mel-spectrum. The triangle filters are distributed on Mel-scale among upper and lower frequency bands for producing energies of filter banks. Both time and frequency domains are obtainable in filter banks which is generally utilized in frequency domain for MFCC calculation. The Mel approximation to physical frequency is calculated by eq. (2),

$$f_{Mel} = 2595 log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

Log: Here, the Mel-filter banks is converted into its respective logarithmic representation.

DCT: By applying DCT into log-converted Mel-frequency coefficients, the cepstral coefficients are produced. The MFCC is calculated by eq. (3),

$$C(i) = \sum_{j=0}^{M-1} log_{10}(S(j)) \cos\left(\frac{\pi n(j-0.5)}{M}\right) \tag{3}$$

Where, M is count of triangle Mel-filters, C is a number of cepstral coefficients and $C(i)$ is ith cepstral coefficient. Totally 13 MFCC features are extracted per frame with first and second-order derivatives which leads total 38 features per frame.

3.4 Speaker Identification and Verification

The extracted features are given as input to identification and verification process. Here, the CNN-LSTM is chosen for identification and verification due to its capability to capture both spatial and temporal features of speech signal. CNN extract local patterns in data whereas LSTM is better at handling sequential data and captures long-term dependencies. The CNN training process includes various types of decisions which includes distinguishing input data, number of convolutional and pooling layer and filter dimension indicates learning rate, number of epochs, dropout and batch size. The convolutional layer, batch normalization, flatten, dense and output layer are explained below:

Convolutional Layers:

The conv2D layer contains 64 filters which receives input shapes $20 \times 5 \times 1$ when kernel size is fixed to 5. The activation function of this layer is Conv2D which uses similar activation function that has similar configuration for kernel size and pooling. It contains 128 filters and input shape of $20 \times 5 \times 64$. The maxpooling2D layer is applied after every two Conv2D layer which minimizes input dimension.

Batch Normalization and Flatten Layer:

The batch normalization quickens training procedure through minimizing problem which is identified as internal covariate shift. It enables few cautious about initialization and enables high learning rate. The previous layer parameters are change at training procedure, less learning rates and parameter initialization. The flatten layer converts conv layer output into 1D array which input to following hidden layer through global average pooling.

Dense Layer and output layer:

The one output and dual dense layers are applied for model preprocessing I which 256 and 512 nodes are considered in dense levels. The Rectified Linear Unit (ReLU) is a nonlinear function which is used at initial two dense layer due to less gradient descent through converting every negative activation into zero. The output layer contains SoftMax activation function which includes numerous nodes as classes for training model. The SoftMax converts logits into possibilities and predict the choice through high possibility. At last, middle norm is applied to output for test sample from every speaker.

The LSTM is utilized to overcome expanding and vanishing gradient problems through memory blocks rather that Recurrent Neural Network (RNN). Data from past layer is remembered and integrated in present through LSTM. It used in combination through input, forget and output gates. Z_k is present input, C_k and c_{k-1} are present and past cell states, H_k and H_{k-1} are previous and current outputs. The LSTM input cell is given in eq. (4-6),

$$I_k = \sigma[W_I \times (H_{k-1}, Z_k) + B_I] \tag{4}$$

$$C'_k = tanh[W_I \times (H_{k-1}, Z_k) + B_I] \tag{5}$$

$$C_k = F_k C_{k-1} + I_k C'_k \tag{6}$$

To define which data fraction is integrated in eq. (4) consumes H_{k-1} and Z_k by sigmoid layer. After transferring by $tanh$ layer through H_{k-1} and Z_k, new data is attained by eq. (5). In eq. (6), C'_k is $tanh$ result, long-term data, C_{k-1} to C_k and current moment data. During this time, input bias matrices of LSTM is B_I and weight matrices are W_I. Through integrating sigmoid and dot product with LSTM forget gate allows data transmission as eq. (7),

$$F_k = \sigma[W_F \times (H_{k-1}, Z_k) + B_F] \tag{7}$$

Where, bias matrices are B_F and weight matrices are W_F which is used to define if forget relevant details from past cell with particular cell gate. The output gate is denoted in eq. (8) and (9), where, bias matrices are B_O and weight matrices are W_O, H_{k-1} and Z_k provides required state for continuing LSTM output unit. Its decision vector contains new data C_k through $tanh$ which retrieved and multiplied to obtain output through Fully Connected (FC) layer.

$$O_k = \sigma[W_O \times (H_{k-1}, Z_k) + B_O] \tag{8}$$

$$H_k = O_k \tanh[C_k] \tag{9}$$

The CNN layers extract local features from input data (spectrograms or MFCCs) and LSTM layer process the sequential data output through CNN and captures temporal dependencies. The integrated CNN and LSTM captures both spatial and temporal features of speech signal lead high performance in mismatched and noisy conditions. It significantly improves reliable and effective speaker verification system in real-world applications which contributes accurate and secure biometric authentication system.

4. Experimental Results

The proposed CNN-LSTM experiments is simulated on Python 3.10 with system requirements of windows 10, i5 processor and 8GB RAM. The proficiency of CNN-LSTM is assessed with evaluation metrics of accuracy, f1-score, recall and precision. The integrated CNN and LSTM captures both spatial and temporal features of speech signal lead high proficiency in mismatched and noisy condition which significantly improves reliable and effective speaker verification system. The mathematical expression of various metrics is given in eq. (10-13),

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

Where,

- TP – True Positive
- TN – True Negative
- FP – False Positive
- FN – False Negative

4.1 Performance Analysis

The proficiency of CNN-LSTM is assessed with evaluation metrics of accuracy, f1-score, recall and precision. Here, the different feature extraction and classifier results are analyzed with various performance metrics. The MFCC proficiency is analyzed with various extraction techniques such as Zero Crossing Rate (ZCR), Constant Q-Transform (CQT) and Linear Predictive Coding Coefficients (LPCC) as shown in table 1. The CNN-LSTM proficiency is analyzed with various classifiers such as RNN, CNN-RNN and Gated Recurrent Unit (GRU) as shown in table 2.

Table 1. Results for different feature extraction

Methods	**Accuracy (%)**	**F1-score (%)**	**Recall (%)**	**Precision (%)**
ZCR	77.83	78.50	78.00	78.00
CQT	83.65	83.00	82.00	82.90
LPCC	90.46	90.15	90.33	90.07
MFCC	**99.09**	**99.09**	**99.09**	**99.09**

In the above table 1, the proficiency of MFCC with various feature extraction techniques are analyzed on collected dataset. The MFCC attained high accuracy 99.09%, f1-score 99.09%, recall 99.09%, and precision 99.09% which shows better performance than state-of-art methods such as ZCR, CQT and LPCC. The MFCC efficiently represent speech signal's power spectrum and widely used for speaker recognition. It captures significant characteristics of human voice which are crucial for distinguishing among speakers.

Table 2. Results for different classifier

Methods	**Accuracy (%)**	**F1-score (%)**	**Recall (%)**	**Precision (%)**
RNN	96.59	96.00	97.00	97.00
CNN-RNN	97.44	98.00	98.00	98.00
GRU	94.95	95.00	95.00	95.00
CNN-LSTM	**99.09**	**99.09**	**99.09**	**99.09**

In the above table 2, the proficiency of CNN-LSTM with various classifiers are analyzed on collected dataset. The CNN-LSTM attained high accuracy 99.09%, f1-score 99.09%, recall

99.09%, and precision 99.09% which shows better performance than state-of-art methods such as RNN, CNN-RNN and GRU. The integrated CNN and LSTM captures both spatial and temporal features of speech signal lead high proficiency in mismatched and noisy condition which significantly improves reliable and effective speaker verification system.

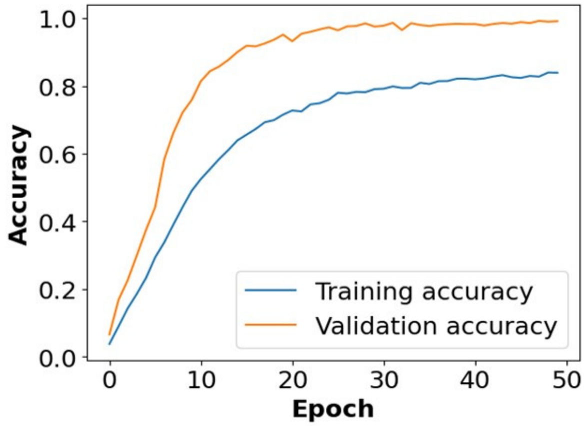

Figure 3. Epoch vs Accuracy

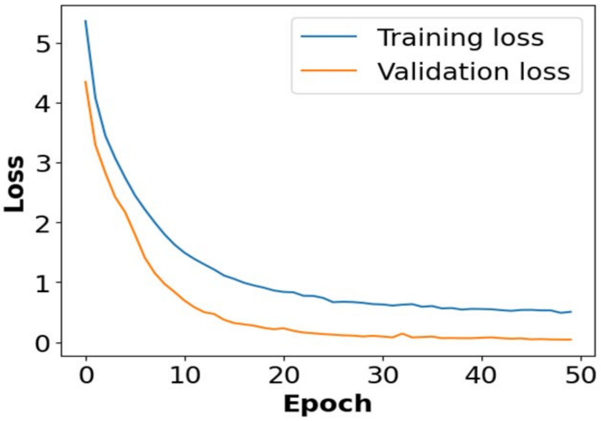

Figure 4. Epoch vs Loss

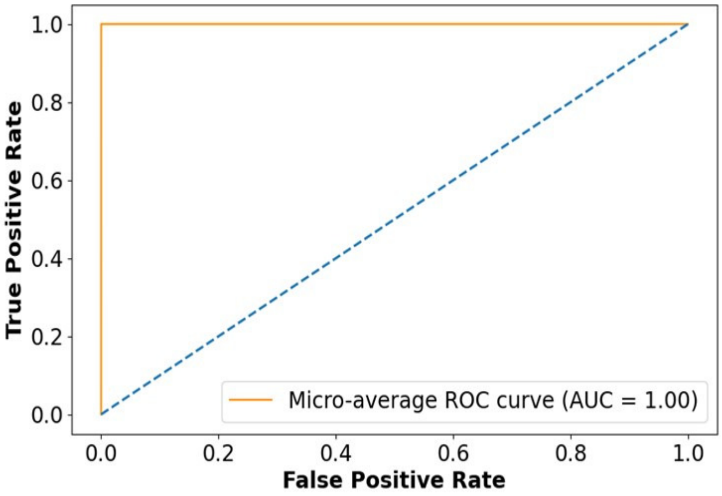

Figure 5. ROC curve

In the above figure 3, 4 and 5 denotes epoch vs accuracy, epoch vs loss and ROC curve for CNN-LSTM. In figure 4, both training and validation accuracy increases in first 20 epochs that denotes that model enhance the performance in both training and validation sets which tackles the overfitting issue. In figure 5, both training and validation loss decreases in first 20 epochs which denotes that model has quick learning and enhance its performance through minimizing error on both training and validation sets. In figure 6, ROC curve is plotted through False Positive Rate (FPR) on x-axis and True Positive Rate (TPR) on y-axis. The ROC curve reaches top left corner which denotes model achieves best separation among positive and negative classes. The Area Under Curve (AUC) is 1.00 that is highest possible value which denotes the model differentiate among classes without error.

4.2 Comparative Analysis

The proposed CNN-LSTM is compared through existing approaches like MLP [20], CNN [21], DNN-ResNet50 [23] and Bi-LSTM [24]. The proficiency of CNN-LSTM is assessed with evaluation metrics of accuracy, f1-score, recall and precision. The CNN-LSTM attained high accuracy 99.09%, f1-score 99.09%, recall 99.09%, and precision 99.09% which shows better proficiency than existing classifier. The CNN and LSTM captures both spatial and temporal features of speech signal lead high performance in mismatched and noisy condition which improves reliable and effective speaker verification system. Table 3 shows the comparative analysis.

Table 3. Comparative Analysis

Methods	**Accuracy (%)**	**F1-score (%)**	**Recall (%)**	**Precision (%)**
MLP [20]	90.2	88.9	88.9	89.1
CNN [21]	82.62	NA	NA	NA
DNN-ResNet50 [23]	98.57	98.74	99.02	98.47
Bi-LSTM [24]	97.94	98.01	98.01	98.02
CNN-LSTM	**99.09**	**99.09**	**99.09**	**99.09**

4.3 Discussion

The existing speech recognition techniques suffers from limitations such as MLP [20] prone to overfitting due to its numerous and complex parameters which provides poor performance. The CNN [21] was unable to capture long-term temporal dependencies which reduces the model performance. DS-GAU [22] was complex to train due to its inability to handle temporal dependencies. DNN-ResNet50 [23] reduces performance and noise robustness with mismatched conditions of training and testing. Bi-LSTM [24] leads overfitting because it learns noisy and irrelevant patterns in training data. To tackle this issue, this research proposed CLNN-LSTM for speaker identification and verification system. The CNN extract local patterns and LSTM captures long-term dependencies and handles sequential data efficiently. The integration of CNN and LSTM leads better performance in noisy and mismatched conditions which enhances the speaker verification system.

5. Conclusion

This research proposed a CNN-LSTM for speaker identification and verification system. The label encoding is used for data preprocessing which converts categorical identities into numerical values. Then, MFCC is used for extracting 13 features which efficiently present speech signal power spectrum and captures human voice features which are crucial for differentiate among speakers. The CNN extract local patterns and LSTM handles sequential data for capturing long-term dependencies. The integration of CNN-LSTM leads high performance in mismatched and noisy conditions which enhances the speaker verification system. The CNN-LSTM achieves higher accuracy 99.09%, f1-score 99.09%, recall 99.09%, and precision 99.09% which shows better performance than existing classifiers. In future, various DL algorithms can be utilized for enhancing speaker identification and verification system.

References

[1] Nirmal, A., Jayaswal, D. and Kachare, P.H., 2024. A Hybrid Bald Eagle-Crow Search Algorithm for Gaussian mixture model optimisation in the speaker verification framework. Decision Analytics Journal, 10, p.100385.

[2] Abdelwahab, K.M., El-atty, S.A., Brisha, A.M. and Abd El-Samie, F.E., 2022. Efficient cancelable speaker identification system based on a hybrid structure of DWT and SVD. International Journal of Speech Technology, 25(1), pp.279-288.

[3] Liu, T., Das, R.K., Lee, K.A. and Li, H., 2022. Neural acoustic-phonetic approach for speaker verification with phonetic attention mask. IEEE Signal Processing Letters, 29, pp.782-786.

[4] Neelima, M. and Prabha, I.S., 2024. Optimized deep network based spoof detection in automatic speaker verification system. Multimedia Tools and Applications, 83(5), pp.13073-13091.

[5] Salim, S., Shahnawazuddin, S. and Ahmad, W., 2023. Automatic speaker verification system for dysarthric speakers using prosodic features and out-of-domain data augmentation. Applied Acoustics, 210, p.109412.

[6] Bharath, K.P. and Kumar, M.R., 2022. Replay spoof detection for speaker verification system using magnitude-phase-instantaneous frequency and energy features. Multimedia Tools and Applications, 81(27), pp.39343-39366.

[7] Pandian, J.A., Thirunavukarasu, R. and Kotei, E., 2024. A Novel Convolutional Neural Network Model for Automatic Speaker Identification from Speech Signals. IEEE Access.

[8] Garain, A., Ray, B., Giampaolo, F., Velasquez, J.D., Singh, P.K. and Sarkar, R., 2022. GRaNN: feature selection with golden ratio-aided neural network for emotion, gender and speaker identification from voice signals. Neural Computing and Applications, 34(17), pp.14463-14486.

[9] Farsiani, S., Izadkhah, H. and Lotfi, S., 2022. An optimum end-to-end text-independent speaker identification system using convolutional neural network. Computers and Electrical Engineering, 100, p.107882.

[10] Nassif, A.B., Shahin, I., Nemmour, N., Hindawi, N. and Elnagar, A., 2023. Emotional Speaker Verification Using Novel Modified Capsule Neural Network. Mathematics, 11(2), p.459.

[11] Gaurav, Bhardwaj, S. and Agarwal, R., 2023. An efficient speaker identification framework based on Mask R-CNN classifier parameter optimized using hosted cuckoo optimization (HCO). Journal of Ambient Intelligence and Humanized Computing, 14(10), pp.13613-13625.

[12] El-Gazar, S., El Shafai, W., El Banby, G.M., Hamed, H.F., Salama, G.M., Abd-Elnaby, M. and Abd El-Samie, F.E., 2022. Cancelable Speaker Identification System Based on Optical-Like Encryption Algorithms. Comput. Syst. Sci. Eng., 43(1), pp.87-102.

[13] Qin, Y., Ren, Q., Mao, Q. and Chen, J., 2023. Multi-branch feature aggregation based on multiple weighting for speaker verification. Computer Speech & Language, 77, p.101426.

[14] Zhao, Y., Togneri, R. and Sreeram, V., 2022. Multi-task learning-based spoofing-robust automatic speaker verification system. Circuits, Systems, and Signal Processing, 41(7), pp.4068-4089.

[15] Jin, R., Ablimit, M. and Hamdulla, A., 2023. Speaker verification based on single channel speech separation. IEEE Access.

[16] Karthikeyan, V., 2022. Adaptive boosted random forest-support vector machine based classification scheme for speaker identification. Applied Soft Computing, 131, p.109826.

[17] Al-Karawi, K.A., 2024. Face mask effects on speaker verification performance in the presence of noise. Multimedia Tools and Applications, 83(2), pp.4811-4824.

[18] Alnuaim, A.A., Zakariah, M., Shashidhar, C., Hatamleh, W.A., Tarazi, H., Shukla, P.K. and Ratna, R., 2022. Speaker gender recognition based on deep neural networks and ResNet50. Wireless Communications and Mobile Computing, 2022, pp.1-13.

[19] Saritha, B., Laskar, M.A., Kirupakaran, A.M., Laskar, R.H., Choudhury, M. and Shome, N., 2024. Deep Learning-Based End-to-End Speaker Identification Using Time–Frequency Representation of Speech Signal. Circuits, Systems, and Signal Processing, 43(3), pp.1839-1861.

[20] Ayvaz, U., Gürüler, H., Khan, F., Ahmed, N., Whangbo, T. and Bobomirzaevich, A.A., 2022. Automatic Speaker Recognition Using Mel-Frequency Cepstral Coefficients Through Machine Learning. Computers, Materials & Continua, 71(3).

[21] Abraham, J.T., Khan, A.N. and Shahina, A., 2023. A deep learning approach for robust speaker identification using chroma energy normalized statistics and mel frequency cepstral coefficients. International Journal of Speech Technology, 26(3), pp.579-587.

[22] Dua, M., Sadhu, A., Jindal, A. and Mehta, R., 2022. A hybrid noise robust model for multireplay attack detection in Automatic speaker verification systems. Biomedical Signal Processing and Control, 74, p.103517.

[23] Tsai, T.H. and Khoa, T.D., 2023. DS-GAU: Dual-sequences gated attention unit architecture for text-independent speaker verification. Machine Learning with Applications, 13, p.100469.

[24] Radha, K. and Bansal, M., 2023. Closed-set automatic speaker identification using multi-scale recurrent networks in non-native children. International Journal of Information Technology, 15(3), pp.1375-1385.

[25] Al-Anzi, F.S., 2022. Improved noise-resilient isolated words speech recognition using piecewise differentiation. Fractals, 30(08), p.2240227.

	Mr.Sukumar B S received the BE degree in Electronics and Communication Engineering (ECE) and M. Tech degree in Digital Electronics and Communication Systems (DECS) from Visvesvaraya Technological University (VTU). Presently pursuing Ph.D degree in AI based Signal processing in VTU, India. Currently he is working as Assistant Professor in the Department of Engineering at C.Byregowda Institute of Technology (CBIT), Kolar, India. His research interests include Signal processing, Artificial Intelligence, Embedded Systems, Electronic Instrumentation, Power electronics and IoT. He is a member of the IEEE, LMISTE, IEAE. He can be contacted at email: sukumar.svm@gmail.com.
	Dr.G N Kodanda Ramaiah received the BE degree in Instrumentation & Technology at Sri J C College of Engineering, Mysore from Mysore University in 1997 and M. Tech degree in Bio Medical Instrumentation at Sri J C College of Engineering, Mysore from Mysore University in 2001, Ph. D degree in speech signal processing from JNTU India. Currently he is working as Professor and Head of the Department of Electronics and Communication Engineering at Kuppam Engineering College, Kuppam, India. His research interests includes Signal processing, Artificial Intelligence, Bio Medical Instrumentation, IOT and Embedded Systems. He is a member of the IETE, LMISTE, MIE. He can be contacted at email: gnk.ramaiah@gmail.com
	Dr. Sarika Raga received the B.E. degree in Electronics and Telecommunication at MGM College of Engineering Nanded from Baba Saheb Ambedkar Marathwada University, Aurangabad in 1996. M.E. degree in power electronics at PDA Collage Engineering Kalaburagi from Gulbarga University in 1999. Ph. D degree in Electronics and Telecommunication from Swami Ramanand Tirth Marathwada University-SGGSIET Nanded in 2012. Currently she is working as Program Coordinator and Associate Professor, Department of ECE, VTU, PG Centre, Bangalore Region, Muddenahalli, Chikkaballapura, India. Her research interests includes Power Electronics, Signal processing, Artificial Intelligence, IOT and Embedded Systems. She is a member of the LMISTE. She can be contacted at email: raga.sarika@gmail.com
	Dr. Lalitha Y. S received the B.E. degree in Electronics and Communication Engineering at PDA college of Engineering, Kalaburagi from Gulbarga University. M.E. degree in power electronics at PDA Collage Engineering Kalaburagi from Gulbarga University in 1999. Ph. D degree in Electronics and Telecommunication from Swami Ramanand Tirth Marathwada University-SGGSIET Nanded in 2012. Ph. D degree in Image Processing from VTU [SDMCET, Dharwad], Belagavi. Currently she is working as professor in Electronics and Communication Department, Don Bosco Institute Of Technology Bangalore. Her research interests includes Image Processing, Artificial Intelligence, Natural Language Processing, Embedded Systems. She is member of the LMISTE, IEEE, IETE, TIE. She can be contacted at email: patil.lalitha12@dbit.co.in.
	Dr. G K Venkatesh received the BE in Electronics and Communication Engineering at BIT Bangalore, india. ME in Computer science Engineering at Dr. MGR University, Chennai, india. Ph.D in Electronics at Jain University, Bangalore, Karnataka, india. Currently he is working as Professor and HOD in Electronics and Communication Engineering. His research interests includes Wireless Communication, Embedded Systems, IOT, Networks, Operating Systems. He served as Vice Chairman and Honorary treasurer at IETE, Bangalore center and also as Honorary Secretary of IMAPS, india for consecutive 2 times.