Speaker Identification Using CNN-LSTM Model on RAVDESS Dataset: A Deep Learning Approach

Suryamritha M, Varshini Balaji, Srinidhi Kannan, Murali K.
Department of Computer Science and Engineering, Department of Mathematics
Amrita School of Computing, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India
bl.en.u4aie21126@bl.students.amrita.edu, bl.en.u4aie21139@bl.students.amrita.edu, bl.en.u4aie21121@bl.students.amrita.edu, k_murali@blr.amrita.edu

***Abstract*—Speaker identification is used for identifying an individual based on their voice. Signal processing and deep neural networks are used for feature extraction. This paper presents a method that combines CNN and LSTM for speaker identification. Multiple models such as GMM, CNN, SVM were compared and CNN+LSTM outperformed with an accuracy of 96.52% and F1 measure of 97%. The CNN+LSTM model combines spatial and temporal information extracted by the CNN and LSTM layers which allows to capture of both local and long-term dependencies in the audio data hence making this model very efficient. The above models were evaluated on the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset. The results highlight their potential for practical applications within speaker identification systems.**

***Index Terms*—CNN, LSTM, RAVDESS, GMM, SVM, speaker identification, deep neural networks**

I. Introduction

The speaker is identified using characteristics of voice [1]. Speaker identification has a lot of applications such as voice recognition, security access, and electronic voice eavesdropping [2]. Voice disguise has become a significant threat in illegal activities and it is important to identify the unknown speaker [3]. Identifying a speaker is done using the speech signals and the extracted features [4]. Most of the voice identification system use the mel-scale frequency cepstrum coefficient (MFCC) as the key vocal feature [5]. Speaker identification plays a major role in audio analysis enabling various practical applications such as voice-controlled systems and forensics. Traditional methods are not very efficient and mostly rely on handcrafted features and statistical models which require manual efforts and lacks the adaptability to diverse datasets. Deep learning methods have emerged as a powerful tools for learning intricate patterns from complex data.

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset has high quality recordings of actors that give various emotions such as calm, happy, sad, angry, fearful, surprise and disgust expressions. Each expression has two levels of emotion intensities (normal, strong), with an additional neutral expression. This diverse collection enhances the performance and reliability of speaker identification systems in various real-world scenarios.

The motivation behind this effort is to enhance the speaker identification systems. The fusion of CNNs and LSTMs results in an excellent performance in predicting the speaker. CNNs extract hierarchical features from raw data. MFCCs are widely used feature extraction methods in audio processing, providing informative representations of speech signals. Incorporating LSTM layers enables the system to capture long-term contextual information important for accurate speaker identification. This model achieves superior performance compared to other models promising improved accuracy and robustness in speaker recognition

This paper has compared various models such as SVM, GMM, CNN, and CNN+LSTM in the RAVDESS dataset. Subsequent sections follow a structured format: Section II presents a literature review, Section III details methodology, and contributions, section IV covers results and analysis, and Section V concludes by summarizing findings and suggesting future research directions.

The key contributions of this paper are :

- Contributes by exploring various models, emphasizing their performance in the speaker identification task.
- The paper rigorously analyzes each model's performance for decision making in speaker identification.
- By combining the strengths of CNNs, MFCCs, and LSTMs, this paper aims to contribute to advancing speaker identification systems, paving the way for more effective and adaptable solutions in real-world applications.

II. LITERATURE REVIEW

Advancements in machine learning (ML) and deep learning (DL) are rapidly shaping the fields of speaker recognition (SR) and speech emotion recognition (SER). This collection of papers highlights various approaches to improve accuracy and reliability in SR and SER systems. Researchers continuously strive to enhance performance from traditional methods such as Gaussian Mixture Models (GMMs) to cutting-edge convolutional and recurrent neural networks. Moreover, the incorporation of emotional features alongside traditional speaker traits signifies a shift towards more comprehensive recognition systems. Through thorough evaluation and experimentation,

these studies make substantial contributions to advancing SR and SER technologies.

Nainan et al. [6] conducted a comprehensive analysis of ASR classifiers including GMM, SVM, and 1-D CNN. They introduced a novel method by combining dynamic and static features with feature selection using the Fisher score methodology. Notable accuracy rates were achieved: 1-D CNN (94.77%), SVM (94.51%), and GMM (55.81%). This study advances ASR by promoting the use of dynamic speech aspects and addressing computational complexity through feature selection.

Sefara et al. [7] and Nassif et al. [8] both emphasize the importance of emotional cues in speaker recognition. Sefara et al. developed a speaker recognition system based on emotion using RAVDESS and various machine learning models, highlighting CNN's superiority with a remarkable 92% accuracy. Nassif et al. tackled the challenges of brief utterances, introducing a CapsNet-based model achieving 89.85% accuracy across multiple datasets. Additionally, Shahin et al. [9] proposed a CNN-based model for identification of speaker in stressful environments, outperforming traditional classifiers like SVM and MLP with an average accuracy of 81.6% on Emirati-accented speech. These studies collectively demonstrate the effectiveness of deep learning approaches in addressing challenges in speaker recognition, particularly in handling emotional cues and adverse conditions.

Al Hindawi et al. [10] addresses voice masking and its application in illegal operations. They suggest an improved Support Vector Machine (SVM) classifier to improve high-pitched speaker detection. Using three speech datasets, the study demonstrates that the modified SVM outperforms classical classifiers such as KNN, MLP, RBF, NB, and SVM, with an average performance of 93.95% for disguised voices. Under difficult circumstances, the speaker recognition accuracy is improved by this creative technique.

Hamsa et al. [11] tackle real-time speech processing challenges amidst noise and emotional influences. They propose an end-to-end framework using pre-trained DNN masks and voice VGG to identify auditory events accurately. Their method excels in speaker recognition in adverse conditions, exceeding recent works on speech data based on emotions in English and Arabic. The model provides an efficient solution for voice recognition in difficult situations, achieving identifying speaker rates of 85.2%, 87.0%, and 86.6% for RAVDESS, SUSAS, and ESD, respectively.

Jabnoun et al. [12] present a novel strategy that combines standard speaker qualities with emotional characteristics to improve speaker identification. With the use of a triplet loss model, it extracts speaker ID and emotional information independently. Using a Triplet Loss Function, it integrates Voice Activity Detection (VAD), speaker segmentation, and speaker embedding into a composite speaker recognition model. Combining mood and speaker attributes results in a significant improvement in accuracy, from 72% to 75%, according to extensive examination. Precision metrics, both in macro and weighted averaging, are notably improved with the inclusion of emotional features. While the RAVDESS dataset ensures empirical rigor, challenges arise from merging disparate datasets due to recording environment and equipment variations.

Al-Dulaimi et al. [13] highlight the rising interest in speaker recognition systems, often tested with datasets like TIMIT and RAVDESS. GMMs are common in speaker recognition, while emotional speaker recognition often involves integrating ML and DNN models. TIMIT is a standard benchmark for speaker recognition, and RAVDESS offers emotional recordings for assessment. Recent studies favor DNN models over ML, achieving higher accuracies (92% vs. 88%), signaling a shift to advanced neural network architectures for improved recognition rates.

Vimal et al. [14] delves into MFCC-based audio classification for emotion detection in speech, leveraging ML methods. Using RAVDESS, it extracts features like MFCC and speech signal energy. These features train models employing DT, RF, and SVM algorithms. The research showcases the effectiveness of the RF algorithm, achieving an impressive 88.54% accuracy in emotion classification. This research contributes valuable insights into emotion detection in speech, emphasizing the efficacy of MFCC-based features and ML algorithms in such endeavors.

Kumar et al. [15] conducted a study on speech based emotion recognition, employing CNN-LSTM and Vision Transformer models. They emphasized emotion recognition's importance in enhancing human-machine interaction. Leveraging the EMO-DB dataset, they explored attention-based DL techniques. The CNN-LSTM architecture with an accuracy of 88.50%, while the Vision Transformer model got 85.36%. This research significantly advances emotion recognition technology and its applications in human-computer interaction.

Shraddha et al. [16] provides a comprehensive exploration of child speech recognition utilizing state-of-the-art ASR models, filling a significant gap in research. It focuses on a dataset comprising recordings from 11 children aged 6 to 11, totaling 5,180 utterances, and conducts experiments on 10 hours of child data. Diverse neural ASR models such as Jasper, Wav2Vec, CRDNN, Hubert, and Conformer are investigated, with reported word error rates serving as evaluation metrics. The study assesses model performance across mixed recordings, long recordings (over 5s), and short recordings (under 5s), employing the word error rate metric. Through this rigorous analysis, the paper offers valuable insights into the challenges and opportunities in child speech recognition, paving the way for further advancements in this domain.

Prasanna et al. [17] compare five Speech Emotion Recognition (SER) models using the MELD dataset. Accuracy of the models are: Time Distributed CNN with LSTM (90%), CNN (73%), BiLSTM with attention (75%), CNN + BiLSTM with attention (70%), and Time Distributed CNN with BiLSTM performs the best among the other models.

Kumaran et al. [18] presents a novel approach for Speech Emotion Recognition (SER) employing MFCC and GFCC with a Deep C-RNN. Evaluating on the RAVDESS, the study achieves over 80% accuracy, surpassing previous methods. By

integrating Mel and Gammatone filters in the convolutional layers, the proposed Deep C-RNN model demonstrates superior performance, evidenced by both accuracy and lower loss metrics. Using the RAVDESS dataset highlights how reliable the method is and how it might be used in practical situations.

Terraf et al. [19] emphasizes the significance of varied datasets and sophisticated approaches in speaker identification studies. Cutting-edge techniques like TCEF and MTCEFMTCEF outperform traditional feature extraction techniques and produce encouraging results. TCEF attains significant gains in accuracy, using LSTM on the GRID-NR dataset yielding an 80.88% accuracy rate. Comparably, in the RAVDESS-NR dataset, MTCEFMTCEF attains 83.24% accuracy, demonstrating notable improvements over conventional MFCC features. These results highlight the value of novel approaches in raising accuracy rates on various datasets and analytic levels.

Speaker recognition is studied by Gade et al. [20], who emphasize novel approaches and a variety of datasets. The Improved Biogeography-Based Optimization Algorithm (IBOA) and a number of feature extraction methods (MFCCT, MFCC, GFCC, LPC) are used in conjunction with hybrid LSTM networks. For speaker classification, models such as DNN, DCRNN, GRaNN, PNN, and LSTM are used. Outperforming previous models, the IBOA-based hybrid LSTM network achieves recognition accuracies ranging from 80% to 91.60% across several databases. These findings demonstrate how crucial sophisticated techniques and large-scale datasets are to the advancement of speech recognition.

To sum up, the examined research demonstrate the variety of approaches used in speech emotion recognition (SER) and speaker recognition (SR). Scholars investigate several techniques, ranging from traditional classifiers like GMMs to state-of-the-art DL architectures like Recurrent Neural Networks (RNNs) and CNNs. A comprehensive approach to recognition systems is represented by the incorporation of emotional characteristics alongside conventional speaker qualities. All things considered, these investigations greatly expand SR and SER technologies, offering insightful information and establishing the foundation for further advancement in these fields.

III. Methodology

The Methodology for speaker identification is depicted in Fig1.

A. Dataset Description

The RAVDESS dataset comprises 1440 audio files performed by 24 professional actors, evenly split between genders. Each actor recorded 60 trials, expressing emotions like calm, happy, sad, angry, fearful, surprise, and disgust. These audio files are labelled systematically using a 7-part numerical identifier, indicating modality, vocal channel, emotion, intensity, statement/context, repetition, and actor identity.

B. Data Collection and Preprocessing

The RAVDESS dataset offers a wealth of material for speaker identification studies, featuring recordings of actors expressing a wide range of emotions. To get the data ready for analysis, we start with some initial prep work. We use tools like LabelEncoder and to_categorical functions, to convert the labels into numbers that the model can understand.

C. Feature Extraction

In the process of feature extraction, MFCCs are very essential. They are the building blocks that come from the audio and are used as the main features for training models. They basically capture the sound's short-term power spectrum, which is really important for identifying different speakers. Getting MFCCs is a big deal because it helps us create models that can pick up on the unique things about how someone sounds. Acquiring MFCCs is an important step to build the

D. Data Splitting

The dataset is divided into training and testing sets in an 80:20 ratio. The majority of the data was used to train while a small portion was reserved to evaluate on unseen data to figure out whether the model can work efficiently at any situation accurately.

E. Proposed Methodology

The architecture presented in Figure 2 is a feature fusion architecture for speaker identification based on CNNs and RNNs processing audio features. The input shape of the model is (13, max_length, 1), meaning the MFCC features obtained over time and on a single channel. The CNN branch begins with a Convolution two dimension layer with 64 filters applying a 3×3 kernel and ReLU activation, then batch normalization to reduce variance affecting training. Max pooling has the effect of shrinking the size of the feature map, thus decreasing the number of computations required and reducing the chances of overfitting. This process is followed by another Conv2D layer but with 128 filters. The output is then permuted and reshaped to match the input expected by the LSTM layer.

The reshaped output is passed through an LSTM layer of 64 units to capture temporal information in the audio signal as it is sequential data like speech. The next layer is a dropout layer to reduce overfitting during training by temporarily eliminating some neurons. The LSTM output is then flattened. Simultaneously, another branch takes MFCC features through global average pooling to extract spectral features necessary for speakers' identification. The outputs from the CNN and MFCCs are combined by concatenating the features obtained from the two branches.

The combined feature set using completely connected layer with 512 units and ReLU activation merges and enhances the features. Finally, an output layer with softmax activation function identifies the speaker and has as many nodes as the number of speakers in the dataset. The model is trained using categorical cross-entropy loss and the Adam optimizer, with accuracy as a metric. It is trained for 50 epochs with a batch size of 32, and validation procedures are used to monitor the training process. The proposed architecture incorporates convolutional layers for feature extraction, LSTM layers for

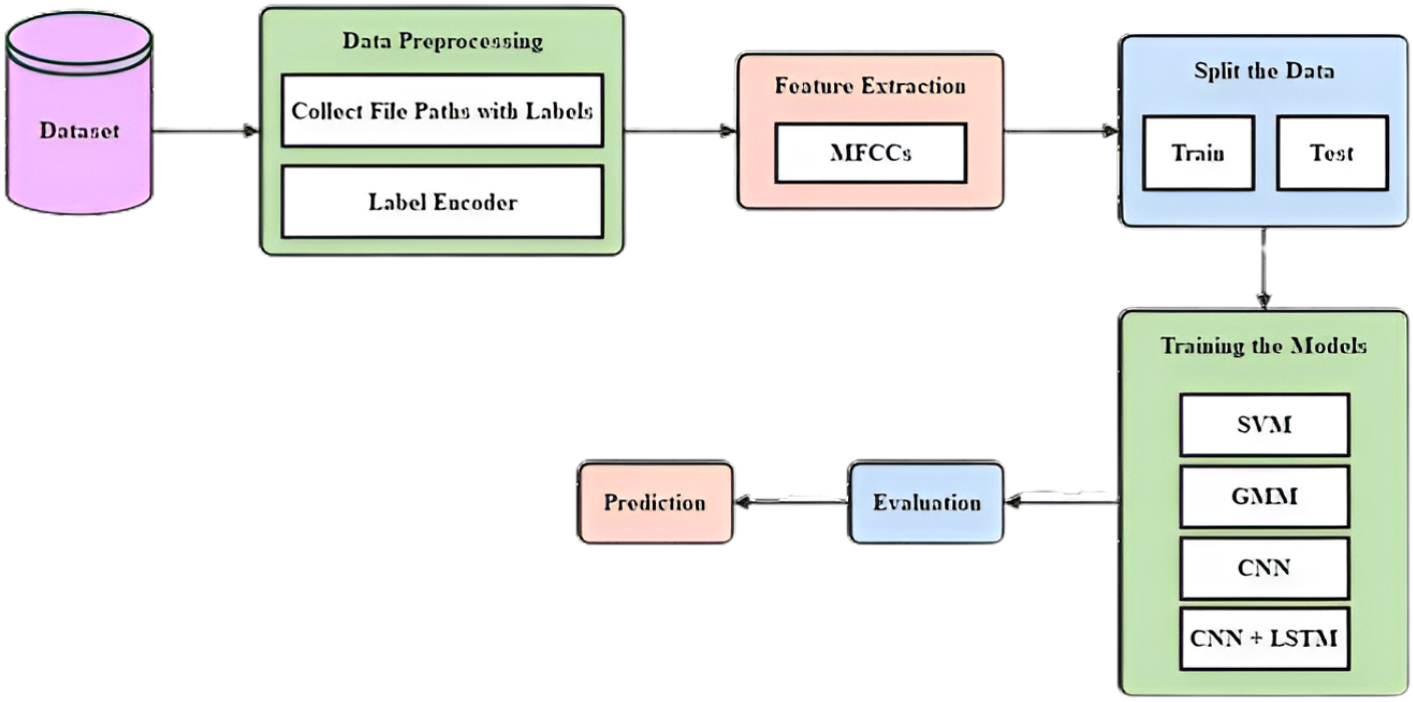

Fig. 1. Methodology for Speaker Identification

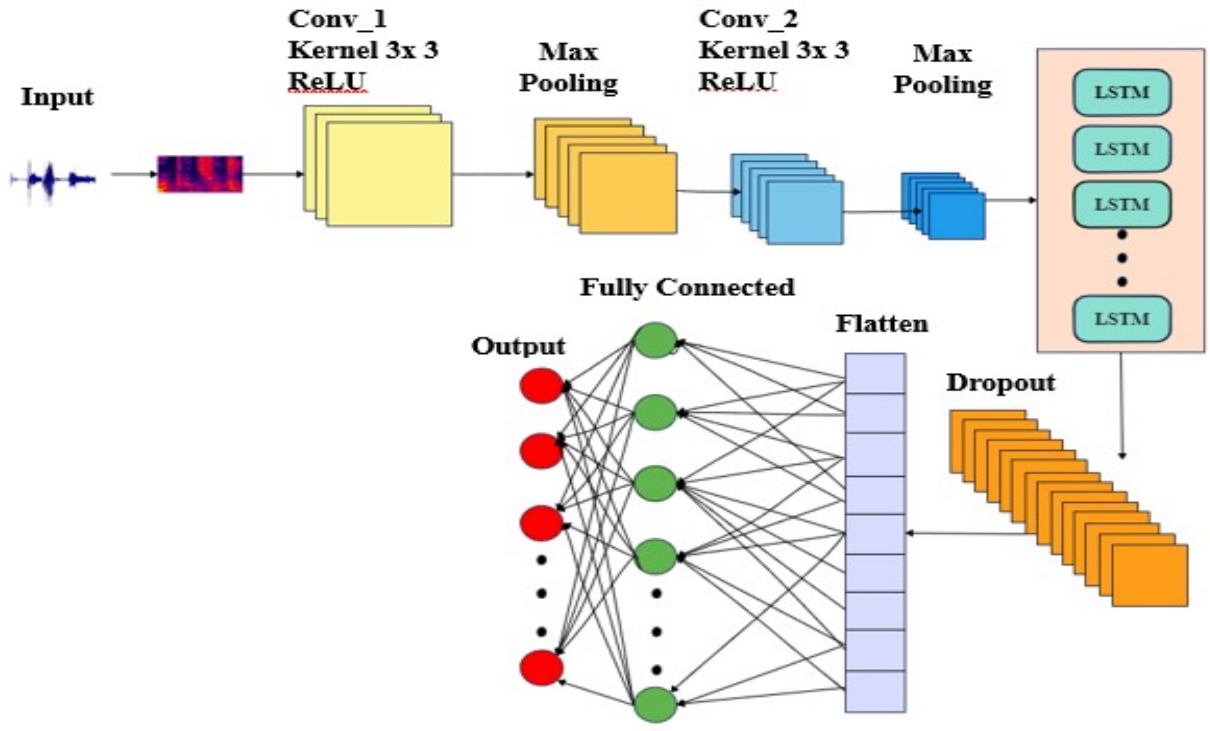

Fig. 2. Architecture of CNN + LSTM Model

temporal analysis, and MFCC features to yield reliable speaker identification.

F. Evaluation

Each model's performance is evaluated using metrics including accuracy, precision, recall, and F1 score. These indicators provide vital insights into the effectiveness and real-world application of the models, allowing for an assessment of their robustness in speaker identification tasks. By testing the models on unseen data helps us understand how well they can adapt to new situations, which is crucial for considering their practical use.

In our comparative study for speaker identification, we explore the effectiveness of four different models: Support Vector Machines (SVMs), CNNs, GMMs, and a hybrid CNN+LSTM model.

1) Support Vector Machine (SVM): SVMs are chosen for their proficiency in handling high-dimensional data like flattened MFCC features. Their ability to represent classes within such spaces makes them well-suited for speaker identification tasks, where distinguishing between different speakers is crucial.

2) Gaussian Mixture Model (GMM): GMMs take a probabilistic approach, modeling the probability distributions of individual speakers based on their MFCC features. This approach helps in better understanding of speaker characteristics, contributing to improved identification accuracy by capturing the variability within speaker data.

3) Convolutional Neural Network (CNN): CNNs known for their capability to learn hierarchical features and excellent in extracting discriminative representations from complex data process MFCCs directly as input.This makes them an attractive option for speaker identification where capturing subtle acoustic patterns is essential.

4) Convolutional Neural Network - Long Short-Term Memory (CNN+LSTM) : The CNN+LSTM model combines the best features of LSTM and CNN architectures. It takes in MFCC features, allowing it to understand the sequence of speech using LSTM and extract spatial information from the audio with CNN. This combination is really helpful because it lets us capture both the order of speech and the spatial details, making it easier to identify speakers accurately.

IV. Results

A variety of models such as GMM, CNN, SVM and CNN+LSTM were trained on RAVDESS for speaker identification. The summarized test results are shown in Table 1.CNN+LSTM achieved the highest F1-score of 96% and accuracy of 97% showcasing its robust performance and the lowest is GMM with an F1-score of 81% and accuracy of 81%.

In the baseline comparison as shown in Table 2, Hamsa et al. achieved 85.2% accuracy using Deep Neural Network mask and speech VGG, and Al-Dulaimi et al. achieved 96.38% on the RAVDESS dataset. Our proposed model excels at 96.52% compared to CNN and LSTM models. The proposed method

TABLE I
Comparison of Models

Sl. No.	Models	Accuracy	Precision	Recall	F1 Score
1	CNN	0.89	0.88	0.88	0.88
2	SVM	0.87	0.88	0.87	0.87
3	GMM	0.81	0.85	0.81	0.81
4	CNN+LSTM	0.97	0.97	0.97	0.96

TABLE II
Baseline Comparison

Authors	**Method**	**Dataset**	**Accuracy**
Hamsa et.al[7]	Deep Neural Network mask and speech VGG	RAVDESS	85.2
Al-Dulaimi et.al[11]	MFCC-CNN	RAVDESS	96.38
proposed method	CNN+LSTM	RAVDESS	96.52

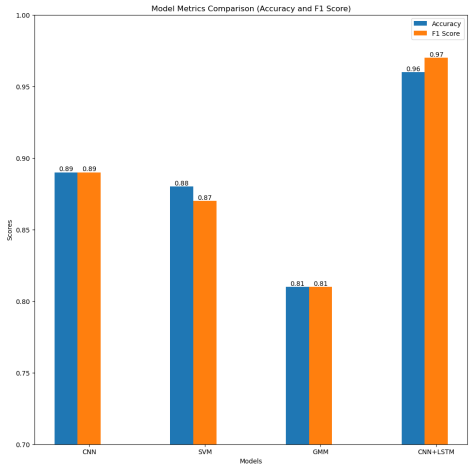

Fig. 3. Model Metrics Comparison

outperforms speaker identification by leveraging CNNs' spatial feature extraction and LSTMs' temporal modeling. This underscores the efficiency of our model in speaker identification.

Figure 3 shows the accuracy and F1 score for all models. Figure 4 shows the CNN+LSTM model's training and validation losses over 50 epochs. The training loss curve depicts the model's loss on the training dataset across epochs, indicating how well the model matches the data. The validation loss, on the other hand, indicates the model's loss on the validation dataset, which includes previously unseen data, and aids in determining if the model is overfitting or underfitting. The graph shows that both training and validation losses are steadily decreasing, showing that the model is learning efficiently and performing well with new data. Low training and validation losses indicate that the model is operating well without overfitting or underfitting.

V. Conclusion and Future Scope

Our study on speaker identification using the RAVDESS dataset has shown some exciting outcomes. The hybrid CNN+LSTM model stood out as the top performer, boasting an impressive F1-score of 96%. This emphasizes the effectiveness of combining both convolutional and recurrent neural

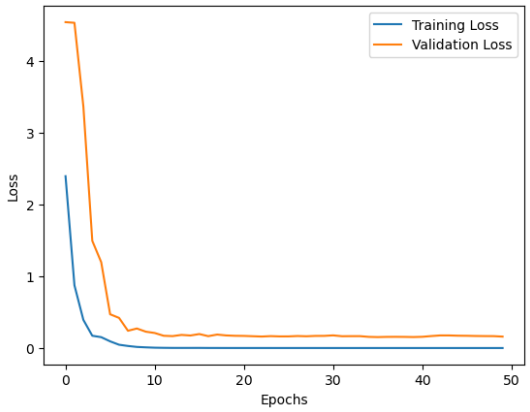

Fig. 4. Training and Validation Loss over Epochs

networks to capture both the temporal and spatial aspects of audio data. These results provide a solid foundation for advancing the accuracy and reliability of speaker identification technology.

Looking ahead, there are several ways we can enhance speaker identification technology. For starters, broadening our dataset to encompass a more diverse range of speakers, spanning various demographics and languages, could be beneficial. Additionally, exploring alternative architectures and features, such as contextual information, may bolster accuracy, particularly in challenging scenarios. Addressing real-world concerns like background noise and diverse recording conditions is also crucial to ensure the practical effectiveness of our technology. Moreover, considerations around privacy and biases must be taken into account when deploying these systems. By continually advancing research and development in these areas, we can improve speaker identification systems both technically and ethically, making them more valuable and socially acceptable.

REFERENCES

[1] Sefara, T.J. and Mokgonyane, T.B., 2020, November. Emotional speaker recognition based on machine and deep learning. In 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC) (pp. 1-8). IEEE.

[2] 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.

[3] Al Hindawi, N.A., Shahin, I. and Nassif, A.B., 2021, March. Speaker identification for disguised voices based on modified SVM classifier. In 2021 18th International Multi-Conference on Systems, Signals & Devices (SSD) (pp. 687-691). IEEE.

[4] Almarshady, N.M., Alashban, A.A. and Alotaibi, Y.A., 2023. Analysis and Investigation of Speaker Identification Problems Using Deep Learning Networks and the YOHO English Speech Dataset. Applied Sciences, 13(17), p.9567.

[5] Ma, H., Zuo, Y., Li, T. and Chen, C.L., 2020. Data-driven decision-support system for speaker identification using e-vector system. Scientific Programming, 2020.

[6] Nainan, S. and Kulkarni, V., 2021. Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN. International Journal of Speech Technology, 24, pp.809-822.

[7] Sefara, T.J. and Mokgonyane, T.B., 2020, November. Emotional speaker recognition based on machine and deep learning. In 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC) (pp. 1-8). IEEE.

[8] Nassif, A.B., Shahin, I., Elnagar, A., Velayudhan, D., Alhudhaif, A. and Polat, K., 2022. Emotional speaker identification using a novel capsule nets model. Expert Systems with Applications, 193, p.116469.

[9] Shahin, I., Nassif, A.B. and Hindawi, N., 2021. Speaker identification in stressful talking environments based on convolutional neural network. International Journal of Speech Technology, 24(4), pp.1055-1066.

[10] Al Hindawi, N.A., Shahin, I. and Nassif, A.B., 2021, March. Speaker identification for disguised voices based on modified SVM classifier. In 2021 18th International Multi-Conference on Systems, Signals Devices (SSD) (pp. 687-691). IEEE.

[11] Hamsa, S., Shahin, I., Iraqi, Y., Damiani, E., Nassif, A.B. and Werghi, N., 2023. Speaker identification from emotional and noisy speech using learned voice segregation and speech VGG. Expert Systems with Applications, 224, p.119871.

[12] Jabnoun, J., Zrigui, A., Slimi, A., Ringeval, F., Schwab, D. and Zrigui, M., 2023, September. Speaker Identification Enhancement Using Emotional Features. In International Conference on Computational Collective Intelligence (pp. 526-539). Cham: Springer Nature Switzerland.

[13] Al-Dulaimi, H.W., Aldhahab, A. and Al Abboodi, H.M., 2022, December. Employing An Efficient Technique with Deep Neural Network for Speaker Identification. In 2022 4th International Conference on Current Research in Engineering and Science Applications (ICCRESA) (pp. 209-214). IEEE.

[14] Vimal, B., Surya, M., Sridhar, V.S. and Ashok, A., 2021, July. Mfcc based audio classification using machine learning. In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-4). IEEE.

[15] Kumar, C.A., Maharana, A.D., Krishnan, S.M., Hanuma, S.S.S., Lal, G.J. and Ravi, V., 2022, December. Speech emotion recognition using CNN-LSTM and vision transformer. In International Conference on Innovations in Bio-Inspired Computing and Applications (pp. 86-97). Cham: Springer Nature Switzerland.

[16] Shraddha, S. and Kumar, S., 2022, June. Child speech recognition on end-to-end neural asr models. In 2022 2nd International Conference on Intelligent Technologies (CONIT) (pp. 1-6). IEEE.

[17] Prasanna, Y.L., Tarakaram, Y., Mounika, Y., Palaniswamy, S. and Vekkot, S., 2022, December. Comparative deep network analysis of speech emotion recognition models using data augmentation. In 2022 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON) (Vol. 2, pp. 185-190). IEEE.

[18] Kumaran, U., Radha Rammohan, S., Nagarajan, S.M. and Prathik, A., 2021. Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN. International Journal of Speech Technology, 24(2), pp.303-314.

[19] Terraf, Y. and Iraqi, Y., 2024. Robust Feature Extraction Using Temporal Context Averaging for Speaker Identification in Diverse Acoustic Environments. IEEE Access.

[20] Gade, V.S.R. and Manickam, S., 2024. Speaker recognition using Improved Butterfly Optimization Algorithm with hybrid Long Short Term Memory network. Multimedia Tools and Applications, pp.1-23.