Multi-Modal Audio Person Identification Using CRNN with BiLSTM and GRU Networks

K. Baby Ramya [1], K. Pavani [2], M.Prasanna [3]

#1 Assistant Professor in the Department of MCA, SRK Institute of Technology, Vijayawada

#2 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.

#3 Student in the Department of MCA, SRK Institute of Technology, Vijayawada

Abstract: This paper proposes an inclusive Person Identification (PID) system that supports both speaking and Non-speaking/Minimal-speaking (NMS) individuals using non-verbal vocalizations. The system utilizes a Convolutional Recurrent Neural Network (CRNN) combined with Supervised Contrastive Learning (SCL) and MFCC feature extraction to effectively process audio signals. To enhance performance, a hybrid architecture integrating Bidirectional LSTM (BiLSTM) and GRU is employed for capturing temporal dependencies. Experimental results demonstrate that the proposed CRNN-BiLSTM-GRU model achieves an accuracy of 96%, outperforming traditional models such as VGG16. The system improves accessibility and provides a robust solution for real-world biometric identification applications.

Index terms - *Person Identification, Non-verbal Vocalization, CRNN, BiLSTM, GRU, MFCC, Supervised Contrastive Learning, Deep Learning, Speaker Recognition, Audio Processing*

1. INTRODUCTION

Person Identification (PID) is a crucial component in various domains such as security systems, healthcare applications, and human-computer interaction. Traditional PID methods include knowledge-based approaches (passwords), token-based systems (smart cards), and biometric-based techniques. Among these, biometric-based identification, especially voice recognition, has gained significant attention due to its convenience, accuracy, and user-friendly nature. Voice-based systems are widely used in applications such as virtual assistants, authentication systems, and access control. However, most existing PID systems rely heavily on speech input, which limits their usability for individuals who are non-speaking or minimal-speaking (NMS).

To address this limitation, this paper proposes an inclusive Person Identification system capable of recognizing both speech and non-verbal vocalizations. The system leverages a Convolutional Recurrent Neural Network (CRNN) integrated with Supervised Contrastive Learning (SCL) to effectively extract and distinguish audio features. Additionally, the incorporation of Bidirectional Long Short-Term

Memory (BiLSTM) and Gated Recurrent Unit (GRU) layers enables the model to capture temporal dependencies in audio signals. By utilizing MFCC features and training on both speech and non-verbal datasets, the proposed system achieves high accuracy and improves accessibility, making it suitable for real-world applications involving diverse user groups.

2. LITERATURE SURVEY

2.1 Person Identification Using Bronchial Breath Sounds Recorded by Mobile Devices:

This study analyzes mobile-sensed BreathPID, which uses breath sounds to identify people. A bespoke breath sound dataset from 21 volunteers is ready for analysis. We train BreathPID's learning models using audio data augmentation (DA) approaches and self-supervised learning (SSL) to overcome spare training data. SSL-based BreathPID models are generated in two phases: solving suggested pretext task(s) without identity information to learn data features, then finetuning the models using labeled data for the downstream task. Several pretense or auxiliary jobs are examined. The primary job for any DA approach is to detect augmentation levels, such as noise introduced to original data samples. When using several DA approaches, the fundamental task is DA type identification. Network architecture, input length variations, and noise resistance are also considered for designing robust BreathPID systems. From the experimental results, SSL-based BreathPID with four DA techniques (noise addition, speed changing, time shifting, and spectrogram masking) yields promising results that are better than SSL-based models using single DA techniques and typical supervised models. The suggested system resists noise and input size variations well. Mobile BreathPID performs as well as or better than stethoscope-sensed BreathPID. The method can be used for mobile device authentication or health monitoring.

2.2 On the use of bronchial breath sounds for person identification:

The sound of turbulent airflow through the glottis, trachea, or large bronchi is bronchial breath sound. It can be used to diagnose respiratory tract and lung disorders and to identify people since it comprises personal physiological information. A neck-mounted stethoscope records bronchial breath sound in this investigation. Mel Frequency Cepstral Coefficients (MFCCs) for each person's bronchial breath sounds are calculated and represented by a stochastic model. By matching the breath sound's MFCCs to each stochastic model, the suggested person identification method identifies the individual who made it. We also use the i-vector technique to improve identification accuracy. Other broad identification systems including support vector machine, random forest, and naive Bayes are used to assess our experimental results' universality. On an 8-person dataset, our investigations indicate that breath sound identification may reach 92% accuracy.

2.3 Stethoscope-Sensed Speech and Breath-Sounds for Person Identification With Sparse Training Data:

In this work, a new biometric termed bronchial breath sound and voice signal obtained by a stethoscope is used to construct a unique person identification (PID) approach. We assess three identification techniques, including support vector machines (SVM), artificial neural networks (ANN), and the i-vector approach, in addition to examining the acoustic properties of breath sounds for PID. This study investigates data

augmentation (DA) strategies that prevent the system training process from overfitting when the training sound data is insufficient, acknowledging the need that the quantity of sound data acquired from each individual should be as modest as feasible. Furthermore, we use feature engineering methods to identify the useful subset of breath sound characteristics for PID. A dataset of 16 people, comprising an equal number of male and female participants, was used for our investigations. Both Artificial Neural Networks and Support Vector Machines in conjunction with feature selection produced encouraging 98% accuracy in the test phase.

2.4 Improved Transcription and Speaker Identification System for Concurrent Speech in Bahasa Indonesia Using Recurrent Neural Network:

One of the most well-known low-resource languages that currently lacks advancements in technology that facilitates communication is Bahasa Indonesia. An enhanced method for producing a transcript and identifying speakers from a concurrent speech in Bahasa Indonesia is presented in this research. The suggested approach may be used in scenarios like distant conferences and online meetings. To identify the speakers in a contemporaneous speech, the system integrates the Reinforced Learning (RL) Model with pitch-aware speech separation. The text transcript is created using a Recurrent Neural Network (RNN) and then enhanced by an external language model and spelling correction model. When compared to alternative approaches, the suggested system was able to identify up to five speakers with varying degrees of confidence and produce higher-quality transcripts for each of them. With an average Word Error Rate (WER) of 16.59% for two speakers, 26.72% for three speakers, and 31.50% for four speakers, the results demonstrate that the suggested approach outperforms the baseline method even in the single-speaker scenario.

2.5 Low Power Speaker Identification by Integrated Clustering and Gaussian Mixture Model Scoring:

This letter describes a new low-power digital CMOS architecture that combines Gaussian mixture model (GMM) scoring with k-means clustering for speaker identification (SI). We demonstrate that k-means clustering at the front-end minimizes downstream processing without compromising SI accuracy by reducing the dimensionality of speech data. To reduce the overhead of the clustering layer (CL), the implementation of a cluster generator with innovative distance calculation and online centroid update datapaths is presented. Among 10 speakers, the integrated design achieves 6× less energy than the traditional method for SI.

3. METHODOLOGY

i) Proposed Work:

The proposed system introduces an inclusive Person Identification (PID) framework designed to recognize both speaking and Non-speaking/Minimal-speaking (NMS) individuals using audio inputs. Unlike traditional speech-dependent systems, this approach processes both verbal and non-verbal vocalizations, ensuring broader accessibility. The system employs Mel Frequency Cepstral Coefficients (MFCC) for effective feature extraction from audio signals, capturing essential frequency-based characteristics

required for accurate identification. A Convolutional Recurrent Neural Network (CRNN) is utilized to learn both spatial and temporal features from the extracted audio representations.

To further enhance performance, the model integrates Supervised Contrastive Learning (SCL) for improved feature discrimination between individuals. Additionally, a hybrid architecture combining Bidirectional Long Short-Term Memory (BiLSTM) and Gated Recurrent Unit (GRU) layers is implemented to capture long-term dependencies and optimize computational efficiency. The system is trained using both the ReCANVo dataset for non-verbal vocalizations and a speaker recognition dataset for speech input. This hybrid CRNN-BiLSTM-GRU model significantly improves identification accuracy, achieving up to 96%, and provides a robust, scalable, and inclusive solution for real-world biometric applications.

ii) System Architecture:

The system architecture begins with the collection of audio data from both speaking and Non-speaking/Minimal-speaking (NMS) individuals. The dataset is passed through a preprocessing stage, which includes visualization and shuffling to ensure balanced and unbiased data distribution. The processed data is then split into training and testing sets. Feature extraction is performed using MFCC, which converts audio signals into meaningful representations suitable for deep learning models. This structured pipeline ensures that both speech and non-verbal vocalizations are effectively prepared for model training and evaluation.

The core architecture involves training multiple models, including the baseline VGG16 and the proposed CRNN-BiLSTM with Supervised Contrastive Learning (SCL), along with its extended version CRNN-BiLSTM-GRU. After training, the models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score. The best-performing trained model is then used in the deployment phase, where users can upload an audio file for person identification. The system processes the input through the trained network and outputs the predicted identity, ensuring high accuracy and an efficient end-to-end identification workflow.

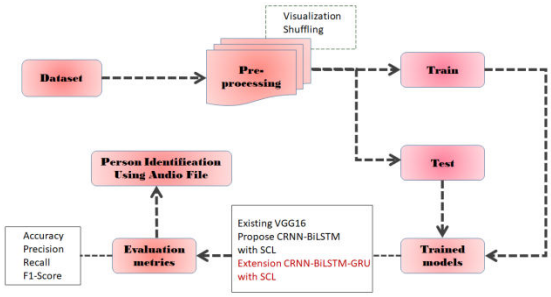

Fig.1. Proposed Architecture

iii) MODULES:

1. Data Loading Module

This module is responsible for importing audio datasets containing both speech and non-verbal vocalizations. It ensures proper organization of data for further processing.

2. Visualization Module

This module visualizes the dataset by displaying the number of audio samples per individual. It helps in understanding data distribution and identifying imbalances.

3. Preprocessing Module (Shuffling & Splitting)

In this module, the dataset is shuffled to remove bias and then split into training and testing sets. This improves model generalization and performance.

4. Feature Extraction Module (MFCC)

This module extracts Mel Frequency Cepstral Coefficients (MFCC) from audio signals. These features represent important frequency characteristics required for accurate identification.

5. Model Training Module

This module trains different models such as VGG16, CRNN-BiLSTM, and CRNN-BiLSTM-GRU with SCL. It learns patterns from both speech and non-verbal audio data.

6. Evaluation Module

This module evaluates model performance using metrics like accuracy, precision, recall, and F1-score. It helps in selecting the best-performing model.

7. Prediction Module

This module takes a new audio file as input and predicts the identity of the person using the trained model.

8. User Interface Module (Flask)

This module provides a web-based interface where users or admins can upload audio files and view identification results easily.

iv) ALGORITHMS:

a. VGG16

VGG16 is a deep Convolutional Neural Network consisting of 16 layers, widely used for feature extraction and classification tasks. In this system, VGG16 is used as a baseline model for person identification by processing MFCC features extracted from audio signals. It captures hierarchical patterns in the data through multiple convolution and pooling layers, enabling basic speaker identification. However, it has limitations in handling temporal dependencies present in audio sequences.

b. CRNN-BiLSTM with SCL

The CRNN-BiLSTM model combines Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks to effectively process sequential audio data. The CNN component extracts spatial features from MFCC representations, while the BiLSTM captures temporal relationships in both forward and backward directions. Additionally, Supervised Contrastive Learning (SCL) is incorporated to improve feature discrimination between individuals by learning more distinct embeddings. This model significantly enhances identification performance compared to traditional approaches.

c. CRNN-BiLSTM-GRU with SCL (Proposed Model)

The proposed model extends the CRNN-BiLSTM architecture by integrating Gated Recurrent Unit (GRU) layers along with BiLSTM. This hybrid approach leverages the strengths of both models, where BiLSTM captures long-term dependencies and GRU improves computational efficiency and faster convergence. The CRNN extracts spatial features, while SCL enhances feature separability for better classification. This combined architecture effectively processes both speech and non-verbal vocalizations, resulting in superior performance. The model achieves the highest accuracy of 96%, making it the most efficient and robust solution for inclusive person identification.

4. EXPERIMENTAL RESULTS

The proposed Person Identification system was evaluated using both speech and non-verbal vocalization datasets. The performance of different models, including VGG16, CRNN-BiLSTM with SCL, and the proposed CRNN-BiLSTM-GRU with SCL, was analyzed. Audio inputs were processed using MFCC feature extraction, and the dataset was divided into training and testing sets (80:20 ratio). Evaluation metrics such as accuracy, precision, recall, and F1-score were used to measure the effectiveness of each model.

The experimental results demonstrate that the proposed hybrid model significantly outperforms the baseline approaches. The VGG16 model achieved an accuracy of 65%, while the CRNN-BiLSTM model improved the performance to 93%. The proposed CRNN-BiLSTM-GRU model achieved the highest accuracy of 96%, along with improved precision, recall, and F1-score. The results confirm that combining BiLSTM and GRU with CRNN and SCL enhances feature learning and classification performance. This makes the system highly reliable and suitable for real-world applications involving both speaking and non-speaking individuals.

Accuracy: The ability of a test to differentiate between healthy and sick instances is a measure of its accuracy. Find the proportion of analysed cases with true positives and true negatives to get a sense of the test's accuracy. Based on the calculations:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Test Accuracy: 0.9895

Precision: The accuracy rate of a classification or number of positive cases is known as precision. Accuracy is determined by applying the following formula:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: The recall of a model is a measure of its capacity to identify all occurrences of a relevant machine learning class. A model's ability to detect class instances is shown by the ratio of correctly predicted positive observations to the total number of positives.

$$Recall = \frac{TP}{(FN + TP)}$$

mAP: One ranking quality statistic is Mean Average Precision (MAP). It takes into account the quantity of pertinent suggestions and where they are on the list. The arithmetic mean of the Average Precision (AP) at K for each user or query is used to compute MAP at K.

$$\boldsymbol{mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k}$$

$$\boldsymbol{AP_k = \text{the AP of class } k}$$

$$\boldsymbol{n = \text{the number of classes}}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic..

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

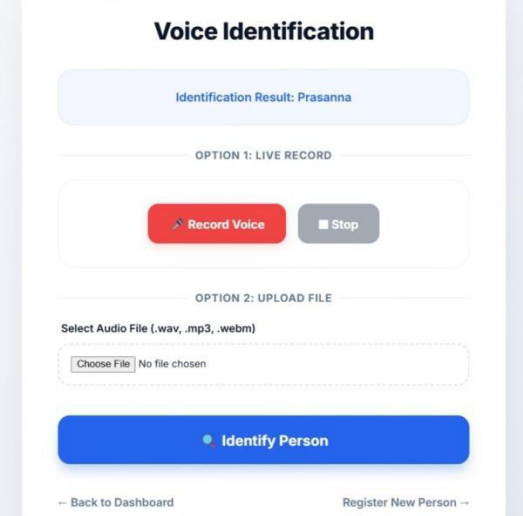

Fig 2:Voice Identification Page

The Voice Identification module analyzes live or uploaded audio samples to identify individuals.The system compares the input voice with stored voice patterns using AI algorithms.It then displays the identification result quickly and accurately.

5. CONCLUSION

This paper presents an inclusive Person Identification (PID) system capable of recognizing both speaking and Non-speaking/Minimal-speaking (NMS) individuals using audio inputs. The proposed approach leverages a Convolutional Recurrent Neural Network (CRNN) integrated with Supervised Contrastive Learning (SCL) and MFCC feature extraction to effectively process both speech and non-verbal vocalizations. By incorporating Bidirectional LSTM (BiLSTM) and GRU layers, the system captures temporal dependencies efficiently, resulting in improved feature learning and classification performance.

Experimental results demonstrate that the proposed CRNN-BiLSTM-GRU model achieves a high accuracy of 96%, outperforming traditional models such as VGG16 and standard CRNN architectures. The system enhances accessibility and inclusivity in biometric identification applications, making it suitable for real-world deployment. Overall, this work provides a robust, scalable, and efficient solution for audio-based person identification across diverse user groups.

6. FUTURE SCOPE

The proposed Person Identification system can be further enhanced by incorporating advanced deep learning architectures such as Transformer and Attention-based models, which can improve the ability to capture complex patterns in audio signals. Additionally, applying transfer learning techniques and pre-trained models can reduce training time and improve performance, especially when dealing with limited datasets. Exploring more sophisticated feature extraction methods, such as spectrogram-based deep features or wavelet transforms, can further increase accuracy and robustness.

In the future, the system can be extended for real-time applications using edge computing and IoT devices, enabling faster and more efficient identification in practical environments. Expanding the dataset to include more diverse non-verbal vocalizations and multilingual speech data will improve generalization. Furthermore, integration with healthcare assistive technologies and security systems can make the solution more impactful, providing inclusive and accessible biometric identification for a wider range of users.

REFERENCES

[1] V.-T. Tran, Y.-L. Lin, and W.-H. Tsai, ‘‘Person identification using bronchial breath sounds recorded by mobile devices,’’ IEEE Access, vol. 11, pp. 66122–66134, 2023, doi: 10.1109/ACCESS.2023.3279502.

[2] V.-T. Tran, Y.-C. Lin, and W.-H. Tsai, ‘‘On the use of bronchial breath sounds for person identification,’’ J. Inf. Sci. Eng., vol. 37, no. 1, pp. 219–241, 2021.

[3] V.-T. Tran and W.-H. Tsai, ‘‘Stethoscope-sensed speech and breath-sounds for person identification with sparse training data,’’ IEEE Sensors J., vol. 20, no. 2, pp. 848–859, Jan. 2020.

[4] M. B. Andra and T. Usagawa, ‘‘Improved transcription and speaker identification system for concurrent speech in Bahasa Indonesia using recurrent neural network,’’ IEEE Access, vol. 9, pp. 70758–70774, 2021, doi: 10.1109/ACCESS.2021.3077441.

[5] N. Iliev, A. Gianelli, and A. R. Trivedi, ‘‘Low power speaker identification by integrated clustering and Gaussian mixture model scoring,’’ IEEE Embedded Syst. Lett., vol. 12, no. 1, pp. 9–12, Mar. 2020, doi: 10.1109/LES.2019.2915953.

[6] S. Nakagawa, L. Wang, and S. Ohtsuka, ‘‘Speaker identification and verification by combining MFCC and phase information,’’ IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 4, pp. 1085–1095, May 2012, doi: 10.1109/TASL.2011.2172422.

[7] H. Yakura, K. Watanabe, and M. Goto, ‘‘Self-supervised contrastive learning for singing voices,’’ IEEE/ACM Trans. Audio, Speech, Language Process., vol. 30, pp. 1614–1623, 2022, doi: 10.1109/TASLP.2022.3169627.

[8] X. Zhang, J. Qian, Y. Yu, Y. Sun, and W. Li, ‘‘Singer identification using deep timbre feature learning with KNN-NET,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun. 2021, pp. 3380–3384, doi: 10.1109/ICASSP39728.2021.9413774.

[9] S. Kooshan, H. Fard, and R. M. Toroghi, ‘‘Singer identification by vocal parts detection and singer classification using LSTM neural networks,’’ in Proc. 4th Int. Conf. Pattern Recognit. Image Anal. (IPRIA), Mar. 2019, pp. 246–250, doi: 10.1109/PRIA.2019.8786009. 68966

[10] W.-H. Tsai and H.-P. Lin, ‘‘Background music removal based on cepstrum transformation for popular singer identification,’’ IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 5, pp. 1196–1205, Jul. 2011, doi: 10.1109/TASL.2010.2087752.

Author Profiles

Ms. K. Baby Ramya is working as an Assistant Professor in the Department of MCA at SRK Institute of Technology, Enikepadu, Vijayawada, NTR District. She completed her MCA from Krishna University. She has nearly 3 years of teaching experience at SRK Institute of Technology. Her areas of interest include Machine Learning, Data Science, and Computer Applications.

Mrs. K. Pavani is working as an Assistant and Head of Department of MCA, in SRK Institute of technology in Vijayawada. She completed her MCA and M.Tech in Computer Science. She has 10 years of teaching experience in SRK Institute of technology, Enikepadu, Vijayawada, NTR District. Her areas of interest include AI and ML, etc.

Ms. M. Prasanna is an MCA student in the Department of Computer Applications at SRK Institute of Technology, Enikepadu, Vijayawada, NTR District. She completed her degree in B.Sc.(Statistics) from Sir CR Reddy College for Women Eluru. Her areas of interest are DBMS and Machine Learning.