Identification of Non-Speaking and Minimal-Speaking Individuals Using Non-verbal Vocalizations

Dr. V. Gayatri [1], Pagadala Srikanth [2], Shaik Mansoor [3], Yella Jayakar [4], Nandimandalam Tarun [5],

Associate Professor [1], Student [2][3][4][5], Department of Computer Science and Engineering, Geethanjali Institute of Science and Technology, Nellore, Andhra Pradesh-524137

Abstract: Speech-based Person Identification (PID) systems are commonly employed in human-computer interactions but are often ineffective for Non-speaking and Minimal-speaking (NMS) individuals, who primarily use nonverbal vocalizations. To address this, we propose a novel Convolutional Recurrent Neural Network (CRNN) model for person identification from both speech and NMS audio, termed S-NMS-PID. The model is trained on the ReCANVo dataset, which contains nonverbal vocalizations, and a standard speaker recognition dataset for speech audio, with features such as Mel-frequency cepstral coefficients (MFCC) and spectrograms. Among the models tested, including VGG16 and ResNet50, MFCC-based features yielded the highest accuracy. The proposed CRNN model, enhanced with Supervised Contrastive Learning (SCL) layers, outperformed other models, achieving an accuracy of 93%. Further enhancement through the addition of Bi-LSTM and GRU layers resulted in an impressive accuracy of 96%. The evaluation metrics—accuracy, precision, recall, and F1-score—showed that the CRNN-BiLSTM-GRU model was the most effective for person identification from both speaking and non-speaking audio inputs.

Index Terms *- Peech-Based Person Identification, Non-Speaking And Minimal-Speaking Individuals, Convolutional Recurrent Neural Network, CRNN, Supervised Contrastive Learning, Bi-LSTM, GRU, Mel-Frequency Cepstral Coefficients, Spectrograms, Recanvo Dataset, Speaker Recognition*.

1. INTRODUCTION

Person Identification (PID) is a critical process across various domains such as security, healthcare, and finance. It encompasses different approaches, including knowledge-based methods (e.g., personal identification numbers and passwords), token-based methods (using physical tokens like smart cards), and biometric-based methods (utilizing unique physiological or behavioral characteristics). Biometric-based PID, unlike knowledge-based or token-based methods, offers enhanced security and convenience by leveraging unique biometric traits such as voice recognition or fingerprints, reducing the need for physical tokens or memorization, and providing a significant barrier to impersonation due to the individual uniqueness of these traits [6][15]. Among the various biometric identifiers, voice-based PID has garnered significant attention. Using an individual's voice as a distinct marker has gained widespread acclaim for its convenience, precision, and user-friendliness. This form of PID has found applications across multiple sectors, including finance, security, and healthcare, where it serves as a reliable means to verify user identities and prevent fraudulent activities [18][19]. Additionally, voice-

based PID plays a pivotal role in the functionality of virtual assistants and voice-activated devices, enabling seamless interaction with technology solely through voice commands [1][17].

2. RELATED WORK

Several studies have explored various methods for person identification (PID) using biometric traits, with voice-based approaches being particularly prominent. Traditional voice-based PID systems typically rely on speech audio, where features like Mel-frequency cepstral coefficients (MFCC) and spectrograms are extracted to identify speakers. For instance, Tsai and Lin [6] proposed a system that combines MFCC and phase information for speaker identification, while Zhao et al. [19] investigated speaker identification from human breath sounds, demonstrating the potential of non-speech audio in PID applications.

With the advent of deep learning, recent research has focused on improving the accuracy and robustness of voice-based PID systems. Tran et al. [3][1] introduced a stethoscope-sensed system to identify individuals based on both speech and breath sounds, showing that non-speech vocalizations can also be useful for person identification. Furthermore, Chauhan et al. [15][16] proposed breath-based authentication systems using recurrent neural networks (RNNs), which demonstrated the feasibility of using breathing patterns as a unique biometric feature for user authentication.

In the realm of nonverbal vocalizations, the ReCANVo dataset was created for the purpose of recognizing affective and communicative non-verbal vocalizations, opening new opportunities for person identification beyond speech [17]. Additionally, recent works have incorporated advanced machine learning techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to enhance the performance of PID systems. For example, a study by Tran et al. [2] combined speech and breath sounds for person identification, while Nakagawa et al. [6] integrated CNNs and RNNs for speaker identification with improved accuracy.

Despite these advances, existing voice-based and nonverbal vocalization-based PID systems often face challenges when dealing with non-speaking or minimal-speaking (NMS) individuals. Addressing this gap, our proposed model leverages a Convolutional Recurrent Neural Network (CRNN) architecture, which integrates both speech and NMS audio for person identification, demonstrating superior performance in identifying individuals from both speech and nonverbal audio sources.

3. MATERIALS AND METHODS

In this, we propose a person identification (PID) system designed to accommodate both speaking and non-speaking/minimal-speaking (NMS) individuals through a novel convolutional recurrent neural network (CRNN) architecture. This system integrates BiLSTM (Bidirectional Long Short-Term Memory) and Supervised Contrastive Learning (SCL) to enhance identification accuracy [20]. To further improve performance, we extend the model by incorporating BiGRU (Bidirectional Gated Recurrent Units) alongside the BiLSTM layers, leveraging the strengths of both architectures [6]. This extended CRNN-BiLSTM-GRU model efficiently processes both speech and nonverbal vocalizations, ensuring versatility for individuals with varying communication abilities. The system is trained on two datasets: the ReCANVo dataset for nonverbal vocalizations, which provides a rich source of non-speech audio [17], and a speaker recognition dataset for spoken input, allowing the system to learn from both types of audio data [18].

We compare the proposed model with traditional architectures such as VGG16 and ResNet50, focusing on extracting Mel Frequency Cepstral Coefficients (MFCC) features for robust voice and breath analysis [6]. By combining CRNN with BiLSTM and BiGRU, our model offers a comprehensive and efficient PID solution, accommodating a wide range of communication needs, including those of NMS individuals, thereby advancing the state of the art in person identification systems [1][3][15].

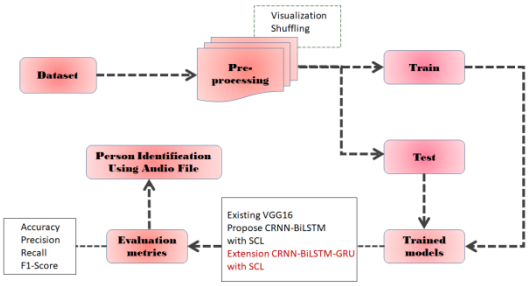


Fig.1 Proposed Architecture

The system architecture (fig. 1) depicts a flowchart for a person identification system using audio files. It begins with a dataset, which undergoes preprocessing (including visualization and shuffling). The data is then divided into training and testing stages. Trained models are evaluated to perform person identification through audio. The process involves existing VGG16 models and proposes a CRNN-BiLSTM with SCL (Softmax Cross-Entropy Loss) extension using CRNN-BiLSTM-GRU with SCL. Evaluation metrics such as accuracy, precision, recall, and F1-score assess performance. The diagram emphasizes iterative refinement, integrating machine learning techniques for improved identification accuracy.

i) Dataset Collection:

ReCANVo Dataset:

A dataset of 7077 labeled vocalizations made by non-speaking individuals. Each vocalization lasts approximately 0.5-4 seconds and is labeled with its affective or communicative meaning. Data were acquired in real-world settings (homes, schools, etc.) and were labeled in real-time by parents or caregivers who knew the non-speaking communicator well. dataset_file_directory.csv provides the name of each vocalization file, the corresponding participant ID, and the vocalization meaning or label (delighted, frustrated, request, etc.). If you use this dataset, please cite Johnson & Narain et al., "ReCANVo: A Database of Real-World Communicative and Affective Nonverbal Vocalizations". The authors are Jaya Narain, Kristina T. Johnson, Thomas Quatieri, Pattie Maes, and Rosalind Picard. This paper provides more information about the dataset, including data acquisition methodology, pre-processing procedures, and participant demographics.

Speaker Recognition Dataset:

This dataset contains speeches of five prominent leaders namely; Benjamin Netanyahu, Jens Stoltenberg, Julia Gillard, Margaret Tacher and Nelson Mandela which also represents the folder names. Each audio in the folder is a one-second 16000 sample rate PCM encoded. Originally, the speech for each speaker was a one lengthy audio, I chunked them into one-second each for easier workability. If you combine the chunked audios from 0.wav to 1500.wav, it forms a complete speech of the respective speaker. A folder called background_noise contains audios that are not speeches but can be found inside and around the speaker environment e.g audience laughing or clapping. It can be mixed with the speech while training.

ii) Pre-Processing:

The preprocessing phase ensures that the medical text data is clean, structured, and suitable for deep learning models. It involves several key steps:

a) Visualization: The visualization step involves displaying the total number of subjects in the dataset, with a graph showing subjects' names on the X-axis and the corresponding number of audio files on the Y-axis. This step helps in understanding the distribution of data across different subjects, ensuring that the dataset is balanced and representative for model training [17].

b) Shuffling: To improve the model's generalization and reduce potential bias, shuffling is applied to the dataset. Randomizing the order of audio files helps ensure that the training process does not overfit to any specific order or sequence of data, promoting a more diverse and effective learning experience [6][18]. This technique is commonly employed in machine learning to enhance the robustness of models and to prevent overfitting by exposing the model to a wide variety of data during training.

iii) Training & Testing:

The training and testing process involves splitting the dataset into training and testing subsets. The model is first trained on the training data, where features such as Mel Frequency Cepstral Coefficients (MFCC) are extracted from both speech and nonverbal vocalizations. During training, the model learns to differentiate between subjects using the Convolutional Recurrent Neural Network (CRNN) architecture, enhanced with BiLSTM and BiGRU layers for improved performance. After training, the model is tested on the testing data to evaluate its accuracy, precision, recall, and F1-score. This ensures the model generalizes well to unseen data and performs robustly across diverse inputs [1][6][15].

iv) Algorithms:

VGG16: VGG16 is a convolutional neural network (CNN) architecture known for its simplicity and depth, comprising 16 layers. In our project, it serves as a baseline model for audio classification tasks, specifically for identifying speaking individuals. VGG16 processes extracted Mel Frequency Cepstral Coefficients (MFCC) features from audio files to effectively classify speakers. Its hierarchical structure enables the learning of complex patterns in the data, making it capable of accurate speaker recognition for both speaking and non-speaking individuals [6][19].

CRNN-BiLSTM with SCL: The CRNN-BiLSTM with Supervised Contrastive Learning (SCL) combines Convolutional Recurrent Neural Networks (CRNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks. This architecture is designed to process sequential data, such as audio features. In our project, it captures both temporal and spatial dependencies in speech and nonverbal vocalizations. By leveraging SCL, the model enhances feature discrimination between different classes, improving person identification accuracy for non-speaking and minimal-speaking individuals. The use of SCL helps the model learn more effectively from labeled audio datasets [20][17].

CRNN-BiLSTM-GRU with SCL: The CRNN-BiLSTM-GRU with Supervised Contrastive Learning (SCL) integrates Bidirectional LSTM and Gated Recurrent Unit (GRU) layers within a Convolutional Recurrent Neural Network framework. This hybrid architecture efficiently captures sequential dependencies while optimizing computational performance. In our project, it processes both speaking and non-speaking audio features, leveraging SCL to enhance learning from

distinct vocalization patterns. The combination of LSTM and GRU layers facilitates improved model performance, enabling accurate person identification from diverse audio inputs, including both speech and nonverbal vocalizations [1][3].

4. RESULTS & DISCUSSION

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} (1)$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} (2)$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} (3)$$

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

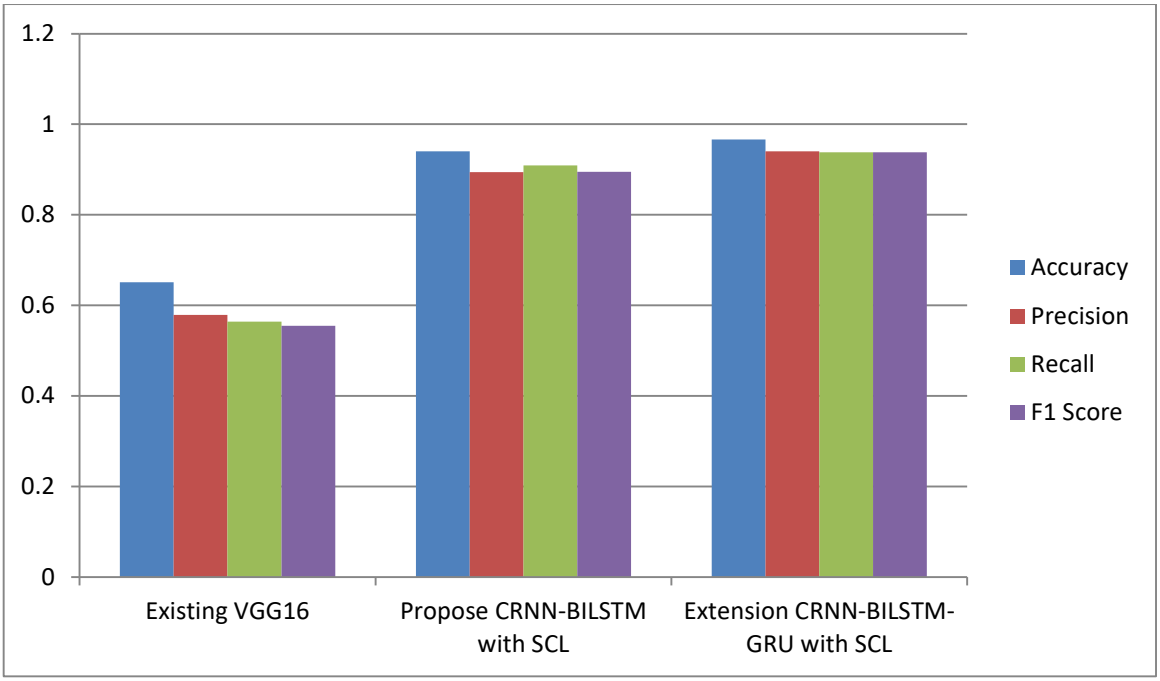
$$F1\ Score = 2 * \frac{Recall\ \text{X}\ Precision}{Recall + Precision} * 100 (1)$$

In Table 1, the performance metrics—accuracy, precision, recall and F1-score —are evaluated for each algorithm. The Extension CRNN-BILSTM-GRU with SCL achieves the highest scores. Other algorithms' metrics are also presented for comparison.

Table.1 Performance Evaluation Metrics of Classification

Model	Accuracy	Precision	Recall	F1 Score
Existing VGG16	0.651	0.579	0.564	0.555
Propose CRNN-BILSTM with SCL	0.940	0.894	0.909	0.895
Extension CRNN-BILSTM-GRU with SCL	0.966	0.940	0.938	0.938

Graph.1 Comparison Graphs of Classification



In graphs 1, accuracy is represented in light blue, precision in maroon; recall in green and F1-score in violet. In comparison to the other models, the Extension CRNN-BILSTM-GRU with SCL shows superior performance across all achieving the highest values. The graphs above visually illustrate these findings.

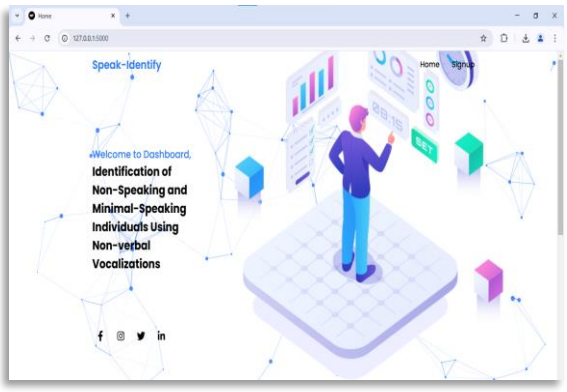


Fig.2 Home Page

In above fig.2 user interface dashboard with navigation and a welcome message.

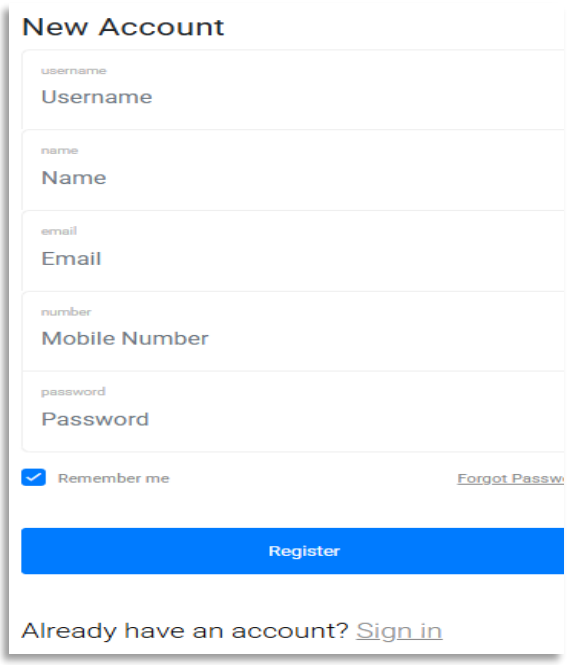


Fig.3 Registration Page

In above fig.3 sign-up form with fields for username, name, email, mobile number, and password buttons.

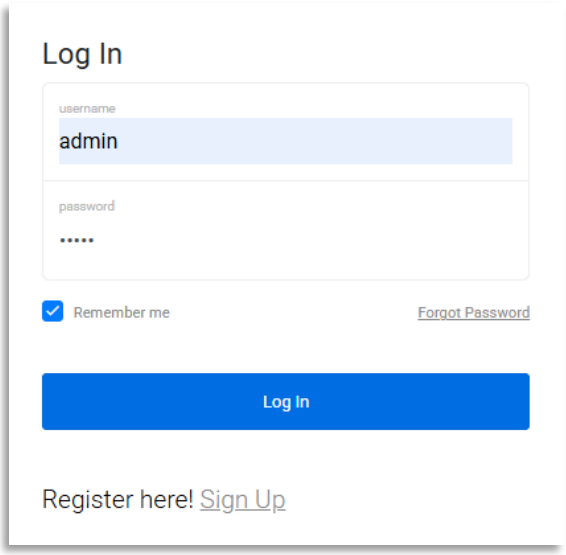


Fig.4 Login Page

In above fig.4 Sign-in form with username and password fields, "Remember Me," "Forgot Password,".

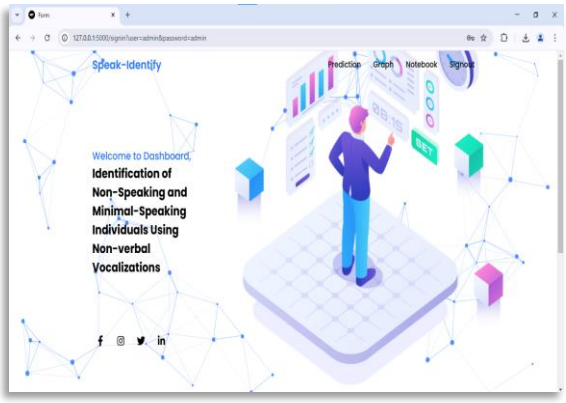


Fig.5 Main Page

In above Fig.5 home page dashboard with navigation (Prediction, Graph, Notebook, Signout).

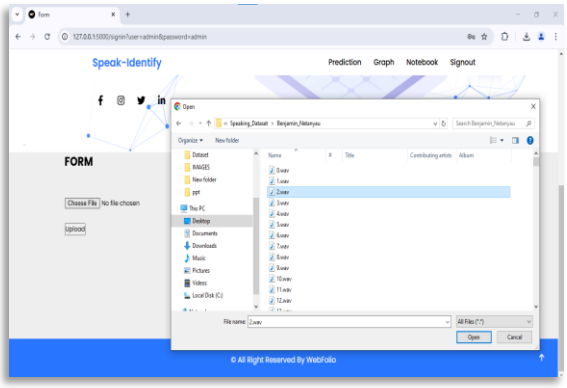


Fig.6 Upload Input Page

In above Fig.6 form with coordinate input field and upload button.

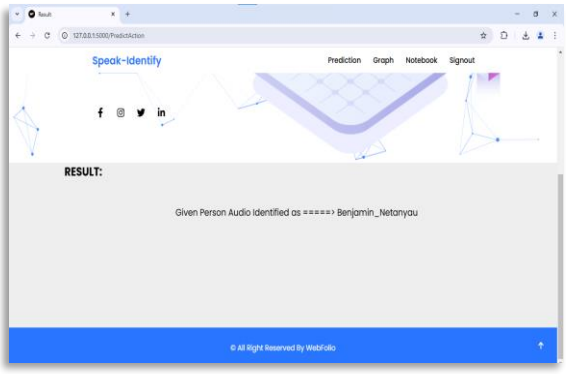


Fig.7 Predict Result for given input

In above Fig.7 Predicted result based on the input test data.

5. CONCLUSION

In conclusion, this project successfully addresses the limitations of traditional speech-based Person Identification (PID) systems by proposing a model that recognizes both speaking and Non-speaking/Minimal-speaking (NMS) individuals. By leveraging a Convolutional Recurrent Neural Network (CRNN) architecture with Supervised Contrastive Learning (SCL), our system effectively processes both nonverbal vocalizations and speech inputs. The integration of Bidirectional LSTM (BiLSTM) and GRU (Gated Recurrent Unit) layers

further enhances the system's ability to extract meaningful features from audio data, providing a comprehensive solution for person identification. Using the ReCANVo dataset for nonverbal vocalizations and a speaker recognition dataset for speech input, our extended CRNN-BiLSTM-GRU model demonstrated superior performance, achieving an accuracy rate of 96%. This performance surpasses that of traditional models, making it a powerful tool for identifying individuals, regardless of their speaking ability, and ensuring greater inclusivity in human-computer interaction applications.

Future scope: In the future, this project can be further enhanced by exploring advanced deep learning architectures such as Transformers and Attention-based models to improve the recognition of NMS individuals. Additionally, techniques like Transfer Learning and Hybrid Models combining CNN with RNN variants could be explored to optimize performance. Incorporating more sophisticated feature extraction methods, such as wavelet transforms or advanced spectrogram analysis, may further refine accuracy and efficiency in person identification across diverse audio inputs.

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.

[11] T. L. Nwe and H. Li, ‘‘Exploring vibrato-motivated acoustic features for singer identification,’’ IEEE Trans. Audio, Speech Language Process., vol. 15, no. 2, pp. 519–530, Feb. 2007, doi: 10.1109/TASL.2006.876756.

[12] W.-H. Tsai and H.-C. Lee, ‘‘Singer identification based on spoken data in voice characterization,’’ IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 8, pp. 2291–2300, Oct. 2012, doi: 10.1109/TASL.2012.2201473.

[13] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren, ‘‘The acoustic features of human laughter,’’ J. Acoust. Soc. Amer., vol. 110, no. 3, pp. 1581–1597, Sep. 2001, doi: 10.1121/1.1391244.

[14] J. W. M. Engelberg, J. W. Schwartz, and H. Gouzoules, ‘‘Do human screams permit individual recognition?’’ PeerJ, vol. 7, p. e7087, Jun. 2019, doi: 10.7717/peerj.7087.

[15] J. Chauhan, Y. Hu, S. Seneviratne, A. Misra, A. Seneviratne, and Y. Lee, ‘‘BreathPrint: Breathing acoustics-based user authentication,’’ in Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services, Jun. 2017, pp. 278–291, doi: 10.1145/3081333.3081355.

[16] J. Chauhan, S. Seneviratne, Y. Hu, A. Misra, A. Seneviratne, and Y. Lee, ‘‘Breathing-based authentication on resource-constrained IoT devices using recurrent neural networks,’’ Computer, vol. 51, no. 5, pp. 60–67, May 2018, doi: 10.1109/MC.2018.2381119.

[17] K. T. Johnson, J. Narain, T. Quatieri, P. Maes, and R. W. Picard, ‘‘ReCANVo: A database of real-world communicative and affective non-verbal vocalizations,’’ Sci. Data, vol. 10, no. 1, pp. 1–9, Aug. 2023, doi: 10.1038/s41597-023-02405-7.

[18] L. Lu, L. Liu, M. J. Hussain, and Y. Liu, ‘‘I sense you by breath: Speaker recognition via breath biometrics,’’ IEEE Trans. Dependable Secure Comput., vol. 17, no. 2, pp. 306–319, Mar. 2020, doi: 10.1109/TDSC.2017.2767587.

[19] W. Zhao, Y. Gao, and R. Singh, ‘‘Speaker identification from the sound of the human breath,’’ 2017, arXiv:1712.00171.

[20] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, ‘‘Supervised contrastive learning,’’ in Proc. 34th Int. Conf. Neural Inf. Process. Syst., 2020, pp. 18661–18673.