Speaker Recognition Using Convolutional Autoencoder in Mismatch Condition with Small Dataset in Noisy Background

Arundhati Niwatkar[1], Yuvraj Kanse[2], and Ajay Kumar Kushwaha[3](✉)

[1] Sivaji University, Kolhapur, Maharashtra, India
[2] Karmaveer Bhaurao Patil College of Engineering, Satara, Maharashtra, India
[3] Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India
akkushwaha@bvucoep.edu.in

Abstract. The objective of this paper is to increase the success rate and accuracy of speaker recognition and identification systems through the proposal of a novel approach. Data augmentation techniques have been employed to enhance a small dataset comprising audio recordings from five speakers, encompassing both male and female voices. The Python programming language is used for data processing. The chosen model is a convolutional autoencoder. In order to convert the speech signal into an image, their respective spectrograms have been used. Consequently, a set of images serves as the input for training the autoencoder. A speaker recognition and identification system are developed using the convolutional autoencoder, a deep learning technique. A comparative analysis is conducted of the results against traditional systems reliant on the MFCC feature extraction technique. The proposed system exhibits a high success rate, indicating its efficacy in accurately recognising and identifying speakers. To account for a "mismatch condition," different time durations of the audio signal are utilised during both the training and testing phases. Through a series of experiments involving various activation and loss functions in permutation and combination, the optimal pair for the small dataset is successfully identified, yielding favorable outcomes. In matched conditions, this system has achieved 92.4% accuracy rate.

Keywords: convolutional autoencoder · deep-learning · speaker recognition · MFCC · mismatch condition

1 Introduction

Over the past few years, research scholars have been making significant progress in the field of speaker recognition and identification [1]. This system aims to identify speakers based on their voices and may be divided into two groups: text-dependent and text-independent [2]. Speaker recognition involves determining which trained speech sample best matches the voice of a speaker, and it serves as a means of verifying or denying a speaker's claimed identity. Several traditional systems have been employed

P. Pareek et al. (Eds.): IC4S 2023, LNICST 536, pp. 318–330, 2024.
https://doi.org/10.1007/978-3-031-48888-7_27

for speaker recognition, such as GMM (Gaussian Mixture Model), i-vectors, and HMM (Hidden Markov Models) [3]. A Gaussian Mixture Model is a probabilistic model that represents the weighted sum of Gaussian mixtures. Historically, the Gaussian Mixture Model has demonstrated great success in creating accurate speaker recognition models. However, modern speaker recognition systems increasingly recommend the use of deep learning approaches. Many researchers are actively exploring this approach to develop precise speaker recognition systems. The convolutional neural network (CNN) has been increasing in prominence among deep learning approaches [4]. Designing a speaker recognition system can involve several challenges and hurdles, such as variability in speech signals, limited training data, computational complexity, and adverse recording conditions. Overcoming these hurdles requires a combination of robust algorithms, large and diverse datasets, careful system design, and continuous improvement based on feedback and evaluation. A new algorithm utilising convolutional autoencoders is proposed in this paper to address the challenge of achieving higher accuracy in speaker recognition. Despite attempting various traditional methods previously, the desired level of accuracy has not been achieved. Recognising the limitations of existing approaches, a novel solution based on a convolutional autoencoder architecture has been proposed in this paper. By leveraging the power of convolutional neural networks and autoencoders, the proposed algorithm aims to overcome the hurdles faced by traditional speaker recognition systems.

This paper begins with a comprehensive literature survey, which provides an overview of existing research and advancements in the field of speaker recognition. This section establishes the background and contextualizes the proposed methodology. Following the literature survey, the paper delves into the methodology, presenting the details of the proposed algorithm based on a convolutional autoencoder architecture for speaker recognition. The algorithm's components, such as the convolutional neural network and autoencoder, are described, along with the specific techniques and approaches employed. Subsequently, the paper moves on to the results and discussion section, where the outcomes of the experiments and evaluations conducted on the dataset are presented. The performance of the proposed algorithm is analyzed, compared to existing methods, and discussed in detail. Any significant findings, limitations, or interesting observations are also explored and discussed. Next, the paper concludes with a concise conclusion section summarizing the key contributions of the research, highlighting the strengths of the proposed algorithm, and discussing potential areas for future improvement and exploration. Finally, the references section lists all the cited sources throughout the paper, ensuring proper attribution and facilitating further reading and research for interested readers.

2 Literature Survey

Speaker recognition systems rely on accurate feature extraction from speech signals to distinguish between different speakers. This paper [5] addresses the need for more robust feature extraction methods that can handle different types of speech signals and noise conditions. Another challenge is to achieve domain robustness in speaker recognition systems. Domain robustness refers to the ability of a system to perform well in different

domains or environments, such as different acoustic conditions or speaking styles [6]. A small neural network architecture could also limit the accuracy of the speaker identification system [7]. Speaker verification systems may perform poorly when faced with speakers or acoustic conditions that are not well represented in the training data. This could be due to limitations in the size or diversity of the available datasets. Even with large amounts of training data, it can be challenging to accurately model the wide range of acoustic variations that can occur between speakers, such as differences in accent, age, or gender. Speaker verification systems may perform poorly in real-world scenarios due to the presence of noise, reverberation, or other environmental factors that can degrade the quality of the speech signal [8]. So, using different CNN architectures, feature extraction techniques, and training methods, one can enhance the model's performance [9]. In this paper, several research gaps are addressed, including the small dataset issue, the search for a good loss function, considerations for different acoustic conditions, and domain robustness, through the utilization of this methodology. An additional step is taken in this study by creating a speaker recognition system using a convolutional autoencoder. Different types of autoencoders, such as vanilla autoencoders and denoising autoencoders, are available. In this experiment, a convolutional autoencoder is employed with a very small dataset collected from five different speakers. Furthermore, all recorded samples are kept without any preprocessing. Each utterance has a duration of 3 s. For training and testing, different texts are used, as this system is text independent. It is important to note that the focus of this paper is on the speaker's speech features, so the language of the training and test data does not matter in this case. The proposed methodology is initially explained in this paper, followed by a discussion of the experimental setup, the results of all the experiments, the conclusion, and the future scope of the proposed model.

The work presented has been focused on developing a new automatic speech recognition (ASR) system based on a sparse auto-encoder neural network architecture inspired by the hunting behavior of Harris hawks [10]. The authors propose a new ASR system that uses a sparse auto-encoder network to learn features from speech signals and recognize speech using a deep neural network (DNN) classifier. The proposed system is assessed using the TIMIT dataset and contrasted with other ASR systems that are already in use. The experimental findings demonstrate that the proposed Harris Hawks Sparse Auto-Encoder Networks (HHSAEN) approach outperforms other traditional and deep learning-based ASR systems in terms of recognition accuracy, achieving state-of-the-art performance. In-depth analysis of the suggested system's learned properties and information on the effectiveness of the suggested approach to voice recognition are also included in the study. This research claims that the challenge lies in expanding the dataset used for training and testing in order to assess its efficacy in more diverse scenarios. This study [11] provides an in-depth examination of deep learning algorithms for voice emotion recognition. The authors explore numerous publicly available databases that contain emotional expressions in speech recordings, as well as the challenges and limits connected with these databases. Several deep learning techniques, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks, are then discussed in relation to voice emotion recognition. The authors describe the architectures and training methods for these models, and they

compare their performance on different databases. The paper also includes a detailed discussion of the pre-processing steps that are necessary to prepare speech data for deep learning models, such as feature extraction and normalization. The authors discuss the potential for utilizing larger and more diverse datasets, improving model accuracy, and developing more robust models that can adapt to different languages and cultures. They also mention the possibility of integrating other types of data, such as physiological signals, to further enhance the accuracy of emotion recognition systems. Overall, the authors suggest that there is still much to be explored and improved upon in this field. This [12] describes a method to modify the accent of non-native speakers to improve their speech recognition accuracy. The authors propose a technique that utilizes neural style transfer to modify the accent of non-native speakers' speech by transferring the style of a reference speaker's speech to the non-native speaker's speech. The authors trained a deep neural network to learn the mapping between the spectrograms of the non-native speaker's speech and the reference speaker's speech. They then applied this network to transform the non-native speaker's speech spectrogram to match the reference speaker's spectrogram while preserving the content of the speech. The resulting modified speech was then used as input to a speech recognition system, and the authors found that this approach improved the recognition accuracy of non-native speakers' speech. The authors evaluated their method on two datasets, and the results showed that their approach outperformed several baseline methods for accent modification. In a paper [13], the authors propose a text-independent speaker identification system based on a deep learning model of a convolutional neural network (CNN). The system aims to identify the speaker of an input speech signal without relying on any specific text or speech content. To achieve this, the authors pre-process the speech signal using Mel-frequency cepstral coefficients (MFCCs) and use them as input to the CNN model. The CNN model is trained using a large dataset of speech signals from multiple speakers, and it learns to extract relevant features from the input speech signals that are specific to each speaker. The authors evaluate the performance of their system using two standard datasets, and they report high accuracy rates, demonstrating the effectiveness of their proposed approach. The proposed system has potential applications in various domains, including security, surveillance, and forensics. It is suggested that modifications be made to the deep learning model to increase the accuracy rate. This paper [14] presents the development and evaluation of a deep learning-based Arabic autoencoder speech recognition system for an electro-larynx device, which is a communication aid used by individuals who have lost their natural voice due to laryngectomy. The proposed system aims to improve the recognition accuracy and usability of the device by addressing the challenges of noisy and limited data and the specific characteristics of electro-larynx speech. Several deep learning models, including convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and autoencoder-based models, are trained and assessed on a dataset of electro-larynx speech recordings. In order to increase the amount of training data, they also used data augmentation techniques. They evaluated the models' performance using a range of metrics, such as accuracy, precision, recall, and F1 score. The outcomes of the experiment demonstrated that the autoencoder-based models performed better than the other models in terms of recognition accuracy and robustness to noise and limited data. On the test set, the suggested system's accuracy of 88.37% outperformed the baseline

system's accuracy of 67.86% by a significant margin. Overall, the application of deep learning-based voice recognition systems in this work offers a promising method for enhancing the usability and efficacy of electro-larynx devices. A laryngectomy patient's quality of life and communication skills may be improved by the suggested system. Further studies suggest exploring more features of the speech signal for better accuracy. In their paper [15], the authors present a comprehensive review of speaker identification techniques using artificial intelligence (AI) and machine learning (ML) methods. The paper discusses various AI techniques, such as neural networks, support vector machines, and deep learning, and how they can be used for speaker identification. The authors also review the challenges and limitations of current speaker identification techniques, including issues related to data pre-processing, feature extraction, and model selection. The paper concludes by discussing potential research directions to address these challenges and improve the accuracy of speaker identification systems. Overall, the work done in this paper provides a valuable overview of the current state of the art in speaker identification using AI and highlights the research challenges that need to be addressed to improve the accuracy and reliability of these systems.

According to previous work, speaker recognition systems have several challenges that can affect their accuracy and reliability. One of the primary challenges is the presence of background noise, which can significantly affect the quality of the audio signal and make it difficult to distinguish between different speakers. Another challenge is speaker variability, which can be caused by differences in speech patterns, accents, and language fluency. The use of voice disguising techniques, such as pitch shifting or speaking in different accents, can also pose a significant challenge to speaker recognition systems. Additionally, speaker recognition systems may encounter challenges in handling large datasets, dealing with impostor attacks, and ensuring the privacy and security of the stored voiceprints. To overcome these challenges, this paper proposes a new system that can enhance the accuracy and robustness of speaker recognition systems.

3 Methodology

Figure 1 shows the proposed model for this work. In this study, a dataset comprising voice samples from five speakers was collected in.wav format. The duration of the samples varied from 3 s to 10 s while maintaining a consistent sampling rate of 16 kHz. Since the system under investigation is a text-independent system, different texts were used for training and testing. This means that the speech samples used for training the model contained diverse content, allowing the system to learn speaker-specific characteristics independent of the spoken text. Similarly, during testing, separate texts were employed to evaluate the system's ability to accurately recognize the speakers without relying on specific textual content. Some studies have proven that even with a small dataset, one can build a successful model [16]. Since the dataset is very small, a data augmentation technique has been employed. Techniques such as time stretching, pitch shifting, noise injection, and speed perturbation can introduce variations and increase the effective size of the dataset. In this work, the time stretching technique is employed to augment the small dataset of speech signals. Time stretching involves altering the duration of the speech signal without changing its pitch. By compressing or expanding the time axis,

variations in the temporal characteristics of the speech are introduced. By applying time stretching to the existing speech samples, new instances of the same speech content are generated, but with different durations. This effectively increases the size of the dataset and provides additional training examples for the speaker recognition system. The data augmentation technique can provide a good amount of data for training the model. Hence, the modified dataset is now ready for further experimental analysis. No preprocessing has been done on the voice samples collected from the speakers. After modifying the database, the next step is to convert all voice samples into spectrograms. Convolutional autoencoders work really well when the inputs are images. Creating spectrograms is a way to represent voice samples in the form of an image. Therefore, all the voice samples are converted into spectrograms.

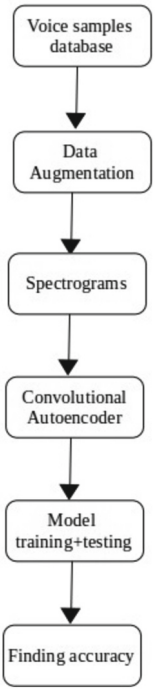

Fig. 1. Speaker Recognition system Framework

Convolutional Autoencoders (CAEs) leverage the power of convolutional operators to capture spatial information effectively. Compared to conventional methods, where convolutional filters are manually designed, CAEs allow the model to discover the ideal filters by minimizing the reconstruction error. Because of their capacity to learn filters, CAEs are the most advanced method for convolutional filter unsupervised learning. CAEs excel at learning concise and useful representations of input data when performing computer vision tasks. CAEs can extract pertinent features from any input data by utilizing the learned filters. These extracted features can then be utilized for various tasks, including classification or any other task that requires a concise representation of

the input. While CAEs are a type of Convolutional Neural network (CNN), there is a fundamental distinction between them. CNNs are typically trained end-to-end, aiming to learn filters and combine features to classify input data. On the other hand, CAEs focus solely on learning filters that can extract features used to reconstruct the input. This differentiation underscores the unique purpose and objective of CAEs in comparison to traditional CNNs. The advantages of using convolutional autoencoders are that they can extract high-level features from raw audio signals, which can result in more accurate speaker recognition compared to traditional feature extraction techniques. They can also effectively filter out noise and other distortions from audio signals, making speaker recognition systems more robust in noisy environments. Speaker recognition systems using CAEs do not require physical contact with the user, making them non-intrusive and convenient to use. The important benefit is that convolutional autoencoders can learn from new data, which makes them adaptable to new speakers and dialects [17, 18]. This adaptability also means that the system can continuously improve its accuracy over time. Hence, by utilizing the convolutional autoencoder in this proposed methodology, the research gaps are overcome.

3.1 Representation of the Database

As shown in the block diagram, all voice samples are converted into spectrograms to be used as inputs for training purposes. Figure 2 depicts the representation of the voice sample as a spectrogram.

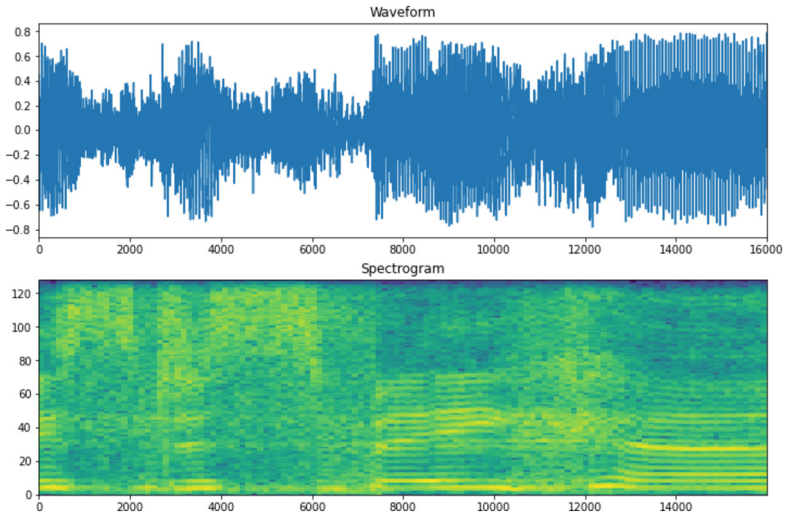

Fig. 2. Representation of voice sample into Spectrogram

3.2 Selection of Activation and Loss Function

Activation functions play a very important role in the system's performance. Its function is to trigger the non-linearity in the cells. The participation of neuron cells will be determined by the activation function. So, in making decisions, it plays a very important role. There are many activation functions, for example, sigmoid. In the proposed model, after experimenting with various combinations, the rectified linear unit (ReLU) is implemented as the activation function, and the mean squared error is chosen as the loss function.

3.3 Details of the Model Used in the Experiment

In this experiment, a convolutional autoencoder is used, which has two parts: an encoder and a decoder. Figure 3 indicates the encoder part, and Fig. 4 indicates the decoder part. The function of pooling layers is the minimization of features. Once the features are minimized, it becomes easy to compute. Here, normalization is also used along with the activation function.

Layer (type)	Output Shape	Param #
resizing_11 (Resizing)	(None, 32, 32, 1)	0
normalization_11 (Normalization)	(None, 32, 32, 1)	3
conv2d_23 (Conv2D)	(None, 30, 30, 32)	320
conv2d_24 (Conv2D)	(None, 28, 28, 64)	18496
conv2d_25 (Conv2D)	(None, 26, 26, 128)	73856

Fig. 3. Convolutional Autoencoder (encoder part)

```
conv2d_transpose_23 (Conv2D  (None, 28, 28, 128)      147584
Transpose)

conv2d_transpose_24 (Conv2D  (None, 30, 30, 64)       73792
Transpose)

conv2d_transpose_25 (Conv2D  (None, 32, 32, 32)       18464
Transpose)

conv2d_transpose_26 (Conv2D  (None, 34, 34, 16)       4624
Transpose)

max_pooling2d_11 (MaxPoolin  (None, 17, 17, 16)       0
g2D)

dropout_22 (Dropout)         (None, 17, 17, 16)       0

flatten_11 (Flatten)         (None, 4624)             0

dense_22 (Dense)             (None, 128)              592000

dropout_23 (Dropout)         (None, 128)              0

dense_23 (Dense)             (None, 5)                645
```

Fig. 4. Convolutional Autoencoder (decoder part)

4 Result and Discussion

The Python programming platform is used for implementing this work. A dataset consisting of 50 voice samples from five distinct speakers has been gathered. The dataset contains utterances ranging from 3 to 10 s, and different texts are used for training and testing purposes. The dataset includes a mixture of languages, but since the expectation is to develop a speaker recognition model based on the speaker's voice features, the language is unlikely to impact the results. Each speaker has a unique characteristic, enabling classification based on the speaker and achieving speaker recognition.

An augmented dataset has been used for training and testing purposes, with the aim of increasing the size of the dataset. The entire dataset has been divided into three sets, namely the training dataset, testing dataset, and validation dataset. In both training and testing, experiments have been conducted under two conditions: the matching condition and the mismatching condition. The matching condition is met when the duration of the training and testing utterances is the same, while in the mismatched condition, the durations are different. The accuracy curve for the matched condition is shown in Fig. 5.

4.1 Result Analysis

In matched conditions, this system has achieved 92.4% accuracy rate. Figure 5 shows the accuracy curve against the number of epochs. It is observed that for a small dataset, the epoch rate should be small. Figure 6 indicates the training loss and validation loss. Every system should be perfectly fitted. Here there is no problem of over- or under-fitting of the model. Figure 6 shows the training loss and validation loss curves in the matched condition (Table 1).

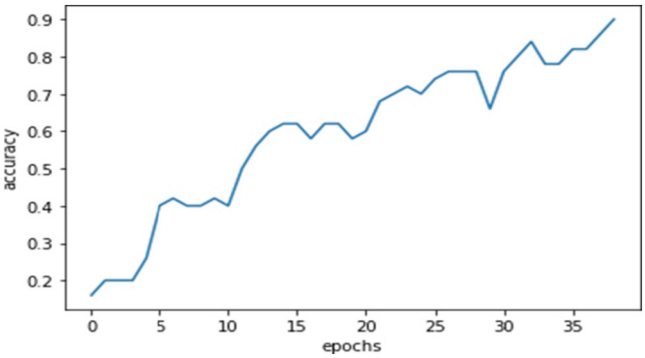

Fig. 5. Accuracy curve for the matched condition.

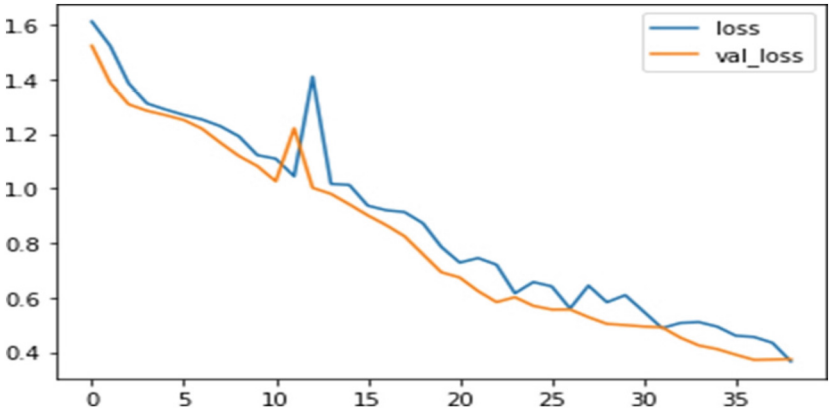

Fig. 6. Loss curve for the matched condition.

Table 1. % Accuracy comparison table of system accuracy in matched and in mismatched condition.

Training sample duration (sec)	Testing sample duration (sec)	
	3	10
3	**92.4**	85.3
10	87.1	**92**

The results indicate that the system has achieved better accuracy in the matched condition, meaning when the test conditions were similar to the training conditions. However, in the mismatched condition, where the test conditions differed significantly from the training conditions, the system's performance decreased. Figure 7 displays the confusion matrix for the matched condition.

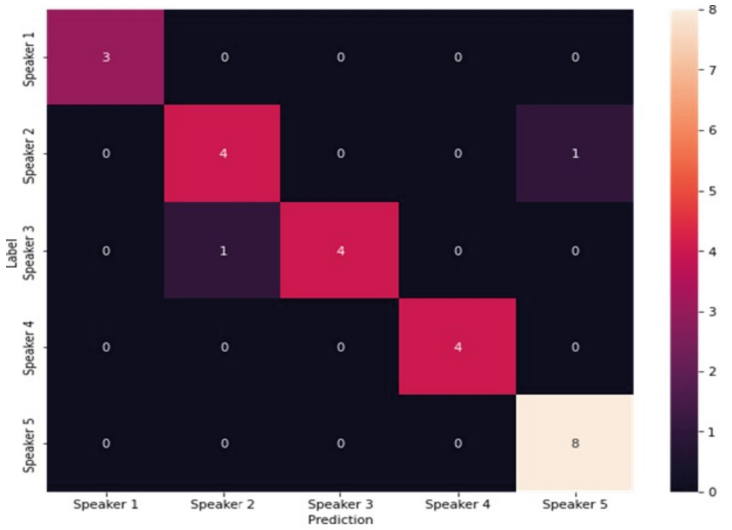

Fig. 7. Confusion matrix for the matched condition.

Table 2 shows the comparison of the proposed model with other methods. As shown in Fig. 7, the confusion matrix shows the performance of the system. It shows the rate of prediction for various labels. Here, five labels have been used. Speaker1, Speaker2, Speaker3, Speaker4, and Speaker5. In the case of matched conditions, the system is performing very well. But in the case of mismatched conditions, its accuracy rate is low. Hence, during training and testing, one can use utterances of the same length.

Table 2. Comparison table of proposed model with other existing methods with same dataset

Method	AUC	CA	F1	Precision
SVM	0.785	0.877	0.839	0.815
Random Forest	0.933	0.853	0.811	0.796
proposed Model	**0.969**	**0.953**	**0.974**	**0.960**

5 Conclusion

In this research paper, a new speaker recognition system is proposed that utilises a convolutional autoencoder. The system has achieved a good success rate under matched conditions for utterances. However, its performance was not satisfactory under mismatched conditions. Various activation functions were experimented with, and it was observed that the ReLU activation function produced better results. The system used raw voice samples without any pre-processing, which made it somewhat resilient to background noise. A comparison was conducted between the proposed system's results and

existing techniques such as SVM and Random Forest using the same dataset. According to Table 2, the system achieved a good accuracy rate. Additionally, parameters like Area under Curve, F1 score, CA, and precision were compared. Previous studies in the related section revealed the use of MFCC features and readily available, clean datasets. However, the novelty of this paper lies in feeding speech signals in the form of images by converting them into spectrograms. Thus, instead of MFCC, the system used spectrograms as a feature of speech signals. Moreover, a specifically collected dataset was used for this research purpose. Based on the experimental results, it has been identified that the main issue lies in the mismatched conditions between the training and testing utterances. Therefore, future researchers should focus on addressing this problem to improve the system's performance. Additionally, since the system was tested with voices containing background noises, it would be beneficial to find a solution to remove these noises, thereby enhancing its performance. Furthermore, it is recommended that researchers explore the use of different types of autoencoders, such as denoising autoencoders and vanilla autoencoders, as well as other feature extraction techniques like pitch, jitter, or LPCC-based feature extraction, to further improve the system's performance.

References

1. La Mura, M., Lamberti, P.: Human-machine interaction personalization: a review on gender and emotion recognition through speech analysis. In: IEEE International Workshop on Metrology for Industry 4.0 & IoT, pp. 319–323 (2020)
2. Shelke, P.P., Wagh, K.P.: Review on aspect based sentiment analysis on social data. In: International Conference on Computing for Sustainable Global Development, pp. 331–336 (2021)
3. Ishak, Z., Rajendran, N., Al-Sanjary, O.I., Razali, N.A.M.: secure biometric lock system for files and applications: a review. In: IEEE International Colloquium on Signal Processing & Its Applications, pp. 23–28 (2020)
4. Hourri, S., Nikolov, N.S., Kharroubi, J.: Convolutional neural network vectors for speaker recognition. Int. J. Speech Technol. **24**, 389–400 (2021)
5. Vaessen, N., Van Leeuwen, D.A.: Fine-Tuning Wav2Vec2 for speaker recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 7967–7971 (2022)
6. Hu, H.R., Song, Y., Liu, Y., Dai, L.R., McLoughlin, I., Liu, L.: Domain robust deep embedding learning for speaker recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7182–7186 (2022)
7. Loina, L.: Speaker identification using small artificial neural network on small dataset. In: International Conference on Smart Systems and Technologies, pp. 141–145 (2022)
8. Lin, W., Mak, M.W.: Robust speaker verification using population-based data augmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7642–7646 (2022)
9. Abdulqader, H.A., Rahman Al-Haddad, S.A., Abdo, S., Abdulghani, A., Natarajan, S.: Hybrid feature extraction MFCC and feature selection CNN for speaker identification using CNN: a comparative study. In: International Conference on Emerging Smart Technologies and Applications, pp. 1–6 (2022)
10. Ali, M.H., et al.: Harris hawks sparse auto-encoder networks for automatic speech recognition system. Appl. Sci. **12**(3), 1091–1095 (2022)
11. Abbaschian, B.J., Sierra-Sosa, D., Elmaghraby, A.: Deep learning techniques for speech emotion recognition, from databases to models. Sensors **21**(4), 1249–1255 (2021)

12. Radzikowski, K., Wang, L., Yoshie, O.: Accent modification for speech recognition of non-native speakers using neural style transfer. EURASIP J. Audio Speech Music Process. **2021**, 11 (2021)
13. Bunrit, S., Inkian, T., Kerdprasop, N., Kerdprasop, K.: Text-independent speaker identification using deep learning model of convolution neural network. Int. J. Mach. Learn. Comput. **9**(2), 143–148 (2019)
14. Zinah J. Mohammed Ameen, Abdul kareem Abdulrahman Kadhim.: Deep learning methods for arabic autoencoder speech recognition system for electro-larynx device. Adv. Hum. - Comput. Interact. **2023**(5), 1–11, (2023)
15. Jahangir, R., Teh, Y.W., Nweke, H.F., Mujtaba, G., Ali Al-Garadi, M., Ali, I.: Speaker identification through artificial intelligence techniques: a comprehensive review and research challenges. Expert Syst. Appl. **171**, 114591 (2021)
16. Jagiasi, R., Ghosalkar, S., Kulal, P., Bharambe, A.: CNN based speaker recognition in language and text-independent small-scale system. In: International conference on IoT in Social, Mobile, Analytics and Cloud, pp. 176–179 (2019)
17. Tirumala, S.S., Shahamiri, S.R.: A deep autoencoder approach for speaker identification. In: International Conference on Signal Processing Systems (2017)
18. Kushwaha, A.K., Khatavkar, S.M., Biradar, D.M., Chougule, P.A.: Depth estimation and navigation route planning for mobile robots based on stereo camera. In: LNICS, Social Informatics and Telecommunications Engineering, vol. 472, pp. 180–191 (2023)