EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Exploring the Impact of Mismatch Conditions, Noisy Backgrounds, and Speaker Health on Convolutional Autoencoder-Based Speaker Recognition System with Limited Dataset

Arundhati Niwatkar [1,*], Yuvraj Kanse[2] and Ajay Kumar Kushwaha[3]

[1] Shivaji 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

Abstract

This paper presents a novel approach to enhance the success rate and accuracy of speaker recognition and identification systems. The methodology involves employing data augmentation techniques to enrich a small dataset with audio recordings from five speakers, covering both male and female voices. Python programming language is utilized for data processing, and a convolutional autoencoder is chosen as the model. Spectrograms are used to convert speech signals into images, serving as input for training the autoencoder. The developed speaker recognition system is compared against traditional systems relying on the MFCC feature extraction technique. In addition to addressing the challenges of a small dataset, the paper explores the impact of a "mismatch condition" by using different time durations of the audio signal during both training and testing phases. Through experiments involving various activation and loss functions, the optimal pair for the small dataset is identified, resulting in a high success rate of 92.4% in matched conditions. Traditionally, Mel-Frequency Cepstral Coefficients (MFCC) have been widely used for this purpose. However, the COVID-19 pandemic has drawn attention to the virus's impact on the human body, particularly on areas relevant to speech, such as the chest, throat, vocal cords, and related regions. COVID-19 symptoms, such as coughing, breathing difficulties, and throat swelling, raise questions about the influence of the virus on MFCC, pitch, jitter, and shimmer features. Therefore, this research aims to investigate and understand the potential effects of COVID-19 on these crucial features, contributing valuable insights to the development of robust speaker recognition systems.

Keywords: MFCC, pitch, jitter, shimmer, convolutional autoencoder

Received on 02 January 2024, accepted on 02 April 2024, published on 09 April 2024

doi: 10.4108/eetsis.5697

*Corresponding author. Email: amehendale@umit.sndt.ac.in

1. Introduction

In recent years, considerable strides have been made by research scholars in advancing the field of speaker recognition and identification [1]. This system, which seeks to identify speakers based on their voices, is commonly categorized into two groups: text-dependent and text-independent [2]. Speaker recognition entails determining which trained speech sample best matches a speaker's voice, serving as a means of verifying or refuting a claimed identity. Traditional systems like Gaussian Mixture Model (GMM), i-vectors, and Hidden Markov Models (HMM) [3] have historically been employed for speaker recognition, with GMM, in particular, demonstrating notable success in creating accurate models. However, contemporary trends increasingly advocate for the adoption of deep learning approaches in speaker recognition systems. Among these, the Convolutional Neural Network (CNN) has gained

prominence [4]. Crafting an effective speaker recognition system poses various challenges, including speech signal variability, limited training data, computational complexity, and adverse recording conditions. Addressing these challenges necessitates a combination of robust algorithms, diverse datasets, meticulous system design, and continuous refinement based on feedback and evaluation. Speaker identification involves determining the identity of a speaker, with successful outcomes relying on natural-sounding recordings captured consistently using a specific instrument or system for identification purposes. The typical workflow for speaker identification encompasses two key stages: training and testing. During training, the system is trained using a dataset comprising diverse voices, and in the testing stage, the trained system is employed to identify voices from a separate set of test voices. To facilitate these tests, the establishment of a database containing various voices is imperative.

Speaker recognition systems are structured around two main stages: feature extraction and speaker classification, utilizing diverse speech characteristics. The features extracted during this process serve as the primary input for speaker identification. The inherent distinctiveness of human voices renders them valuable for various applications. Despite ongoing research in this field, challenges persist, and researchers are motivated to address them. One such challenge involves the impact of health conditions on speech features. To ensure the development of a robust speaker recognition system, it is crucial to employ speech features that remain unaffected by health conditions [5]. This paper specifically delves into the exploration of the impact of COVID-19 on speech features. COVID-19 has been observed to cause infections in the respiratory system, affecting the lungs, throat, vocal cords, and other organs involved in sound production. Consequently, it is pertinent to investigate the effects of this disease on three specific speech signal features: Mel-frequency cepstral coefficients (MFCC), pitch, jitter, and shimmer. Understanding these effects contributes to the broader goal of developing speaker recognition systems that can accommodate and adapt to diverse health conditions, ensuring reliability and accuracy in speaker identification processes. This paper introduces a new algorithm utilizing convolutional autoencoder to tackle the challenge of achieving higher accuracy in speaker recognition. Despite previous attempts with various traditional methods, the desired level of accuracy has remained elusive. Acknowledging the limitations of existing approaches, this paper proposes a novel solution based on the convolutional autoencoder architecture. By harnessing the capabilities of convolutional neural networks and autoencoders, the proposed algorithm aims to surmount the hurdles faced by traditional speaker recognition systems.

2. Objectives

The primary goal of this paper is to identify the optimal feature for constructing a speaker recognition system. To achieve this, speech samples from both healthy and unhealthy conditions of speakers are considered for comparison. The paper proposes an advanced speaker recognition system utilizing spectrogram as a feature, incorporating autoencoder technology. The key focus of this research is to develop a speaker recognition system with a small dataset, evaluating its performance in both matched and mismatched conditions. Remarkably, the paper aims to achieve this without employing any preprocessing on the data, ultimately targeting a speaker recognition system with exceptionally high accuracy.

Over recent years, significant progress has been made in the field of speaker recognition and identification [6]. This process involves determining the identity of a speaker based on their voice, categorized into text-dependent and text-independent approaches [7]. The workflow typically consists of two stages: training, where the system is trained using diverse voice datasets, and testing, where the trained system identifies voices from a separate set of test voices. Establishing a database with varied voices is crucial for effective testing.

Speaker recognition systems comprise feature extraction and speaker classification stages, utilizing diverse speech characteristics as primary inputs. Human voice distinctiveness makes it valuable for various applications, but challenges persist, including the impact of health conditions on speech features [8]. This paper focuses on exploring the effects of COVID-19 on specific speech signal features like Mel-frequency cepstral coefficients (MFCC), pitch, jitter, and shimmer.

Recognizing the limitations of existing approaches, the paper proposes a speaker recognition system using a convolutional autoencoder. This approach aims to address challenges such as small dataset issues, optimal loss functions, considerations for different acoustic conditions, and domain robustness. The convolutional autoencoder is employed with a small dataset from five speakers, emphasizing a text-independent system.

The paper also addresses the challenges faced by speaker recognition systems, including domain robustness, the impact of noise, and limitations in neural network architecture. It highlights the importance of utilizing different CNN architectures, feature extraction techniques, and training methods to enhance system performance [9].

In summary, the proposed methodology focuses on overcoming research gaps related to small datasets, loss functions, acoustic conditions, and domain robustness in speaker recognition. The use of a convolutional autoencoder is a novel approach to address these challenges and enhance the accuracy of speaker recognition systems. The paper details the methodology, experimental setup, results, conclusions, and future scope of the proposed model.

In a related work [10], the focus shifts to developing an automatic speech recognition (ASR) system using a sparse autoencoder neural network inspired by Harris hawks' hunting behavior. This system, named Harris Hawks Sparse Auto-Encoder Networks (HHSAEN), outperforms existing ASR systems in terms of recognition accuracy using the TIMIT dataset. The paper emphasizes the challenge of extending the dataset for evaluating the system's effectiveness in different settings.

Another study [11] explores deep learning algorithms for voice emotion recognition, examining various publicly available databases and discussing challenges and limits associated with these datasets. The authors delve into CNNs, RNNs, and LSTM networks, presenting architecture and training methods while suggesting improvements in dataset diversity and model robustness. In [12], a method to modify the accent of non-native speakers using neural style transfer is proposed to improve speech recognition accuracy. The authors employ a deep neural network to learn the mapping between spectrograms of non-native and reference speakers, enhancing recognition accuracy on two datasets. Furthermore, [13] introduces a text-independent speaker identification system based on a CNN. The system utilizes MFCCs as input, achieving high accuracy rates with potential applications in security, surveillance, and forensics. The authors recommend modifications to the deep learning model for increased accuracy.In the context of communication aids, [14] presents a deep learning-based Arabic autoencoder speech recognition system for electro-larynx devices. The proposed system addresses challenges of noise and limited data, outperforming other models in terms of accuracy and robustness.

Finally, [15] provides a comprehensive review of speaker identification techniques using AI and ML methods, discussing various AI techniques and addressing challenges in data preprocessing, feature extraction, and model selection. The paper concludes by suggesting research directions for improving accuracy and reliability.

In conclusion, the existing literature highlights challenges in speaker recognition systems, including background noise, speaker variability, and dataset-related issues. The proposed system in this paper aims to address these challenges and enhance the accuracy and robustness of speaker recognition systems.

3. Methods

3.1. Mel-frequency Cepstral Coefficients (MFCCs)

Speech signals typically exhibit energy within the 5 KHz range, and their temporal characteristics demonstrate stationarity over short time intervals. To analyse the frequency content, the speech signal is divided into short-duration time slots [16]. The Mel-frequency cepstral coefficients (MFCC) model emulates the human auditory system's frequency perception on a non-linear, logarithmic scale. In a comparative experiment, voice samples were collected from individuals in two conditions: healthy and affected by COVID-19, utilizing various mediums such as telephone recordings, voice messages, and social media videos. The MFCC processing involves six primary steps outlined below [17].

Frame Segmentation: The speech signal is divided into smaller frames of short duration to optimize processing. This allows for individual processing of each frame, as applying Fourier Transform to the entire signal may not yield optimal results.

Windowing: To mitigate spectral leakage and emphasize the central portion of each frame, a windowing technique, like the Hanning window, is applied. This technique modifies the frame's amplitude to reduce unwanted artifacts.

Discrete Fourier Transform (DFT) Calculation: Each windowed frame undergoes DFT application, transforming it from the time domain to the frequency domain.

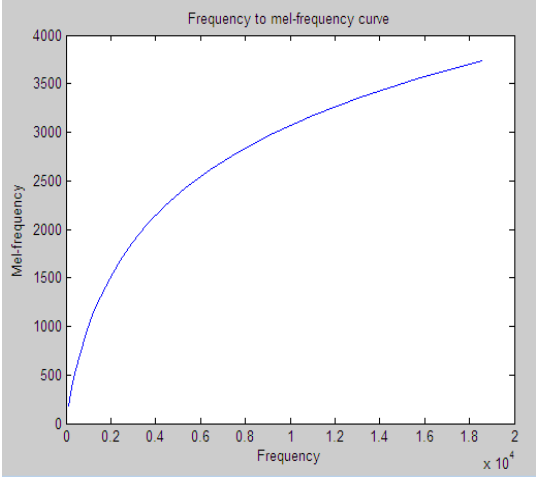

Figure 1. Mapping Frequency to Mel-frequency Scale

Filter Bank Application: To accommodate the human ear's sensitivity to changes on the logarithmic scale, a series of 20-40 triangular filters, referred to as filter banks, are applied. These filters are evenly spaced on the Mel-scale, approximating the human auditory perception of frequency. The utilization of filter banks produces a spectrogram representation of the signal.

Application of Log Scale: The output derived from the filter banks undergoes a logarithmic transformation, converting the filter bank outputs from a linear scale to a log scale. This transformation aligns the representation with the characteristics of the human auditory system. The outcome is a set of log energies that accurately depict the distribution of energy across different filter banks.

3.2. Fundamental frequency (Pitch)

The term "pitch" in speech refers to the perceived frequency or the highness/lowness of a person's voice, primarily determined by the fundamental frequency (Fo) of the vocal cords' vibration during speech production. Pitch is linked to the subjective perception of a person's voice as either high or low, where higher fundamental frequencies correspond to higher pitch, and lower fundamental frequencies correspond to lower pitch. It serves as a crucial element in speech,

conveying information such as emotional expression, gender identification, and linguistic intonation. The fundamental frequency (Fo) is pivotal in speech production, as it is influenced by the vibration rate of the vocal cords when air passes through them. Each individual possesses a distinctive fundamental frequency, shaped by the biological characteristics of their vocal cords. Typically falling within the range of 100 to 400 Hz, females generally have a higher pitch compared to males. In the context of this study, our objective is to investigate the effects of COVID-19 on the fundamental frequency. This examination is conducted using the autocorrelation method for calculation.

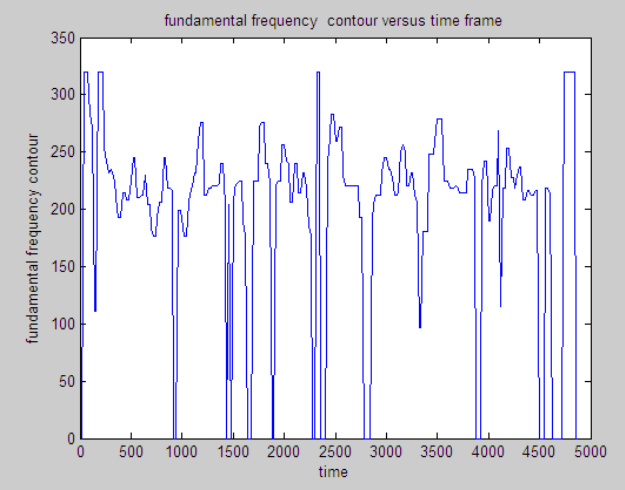

Figure 2. Pitch Vs Time sample graph

3.3. Jitter

In the realm of speech signal analysis, jitter refers to the variation or irregularity in the timing of consecutive periods of a speech waveform. It quantifies perturbations or small deviations in the duration of speech segments, such as fundamental periods or glottal cycles, offering a measure of the instability or irregularity in the vibration of the vocal folds during speech production [18].

3.4. Shimmer

In the domain of speech signal analysis, shimmer refers to variations or fluctuations in the amplitude or intensity of consecutive glottal cycles or vocal fold vibrations. It quantifies perturbations or irregularities in the magnitude of the speech signal waveform, serving as a measure of the instability or variability in the amplitude of the voice signal. Shimmer can perceptually indicate rapid changes in vocal fold vibrations and may be associated with the perceived roughness or hoarseness of a person's voice [19].

In this examination, Discrete Cosine Transform (DCT) is opted to process coefficients acquired from a filter bank. This decision was influenced by the noted high correlation among these coefficients. Specifically, we utilized the DCT to extract Mel-Frequency Cepstral Coefficients (MFCC) from two distinct conditions: healthy and infected. The Euclidean distance between healthy and infected speakers, measured between speech samples from the same speaker but in different conditions, is presented in Table 1. For comparison purposes, focus was on the first 13 MFCC coefficients. The Euclidean distance served as a metric for comparing the MFCC coefficients derived from the two conditions, with a distance of zero indicating no changes in the MFCC coefficients between the healthy and infected conditions. Based on analysis, the results are summarized as follows. Table 1 displays the Euclidean distance between healthy and infected speakers, measuring the distance between speech samples from the same speaker but in two different conditions: healthy and infected.
Additionally, the comparison table for pitch frequencies in infected and non-infected conditions is provided below.

Table 1. Results of MFCC Comparison

Here, I: infected condition, H: Healthy condition

	Speaker1(H)	Speaker 2(H)	Speaker 3(H)	Speaker 4(H)	Speaker 5(H)
Speaker1(I)	0.6675	-	-	-	-
Speaker2(I)	-	0.7797	-	-	-
Speaker3(I)	-	-	0.7845	-	-
Speaker4(I)	-	-	-	0.5346	-
Speaker5(I)	-	-	-	-	0.6290

Table 2. Results of Pitch Comparison

	Pitch in Hz (in H condition)	Pitch in Hz (in I condition)
Speaker1	230.2440	255.669
Speaker2	162.6453	188.7494
Speaker3	268.9532	286.6102
Speaker4	150.7534	179.0453
Speaker5	235.9532	279.0934

Here, I: infected condition, H: Healthy condition

Table 3. Results of Jitter Comparison

	jitter (in H condition in %)	jitter (in I condition in %)
Speaker1	0.24	0.40
Speaker2	0.18	0.39
Speaker3	0.09	0.29
Speaker4	0.17	0.48
Speaker5	0.11	0.35

Here, I: infected condition, H: Healthy condition

Table 2 illustrates the calculation of pitch values for a speaker in both healthy and infected conditions using the autocorrelation method. Clearly, in the infected condition, there is a significant and drastic change in pitch values. Tables 3 and 4 provide jitter and shimmer values in both conditions, revealing notable differences. In the healthy condition, jitter values are low, whereas in the infected condition, these values have increased. The experiment was systematically repeated with five different speakers.

A parallel experiment was conducted to assess the impact of Covid-19 on the shimmer parameter, and the summary is presented in Table 4. The primary objective of this experiment was to identify a robust feature that remains consistent in both healthy and infected conditions. This sought-after feature could then be utilized to construct a resilient speaker recognition system characterized by high accuracy and precision values.

Table 4. Results of shimmer Comparison

	shimmer (in H condition in dB)	shimmer (in I condition in dB)
Speaker1	1.9	3.9
Speaker2	2.1	4.2
Speaker3	3.2	3.8
Speaker4	2.7	5.3
Speaker5	1.8	5.1

Here, I: infected condition, H: Healthy condition

This study focused on crucial aspects of human speech, specifically examining MFCC coefficients, fundamental frequency, jitter, and shimmer due to their significance in speech production. In the initial experiment, it compared the MFCC coefficients of the same speaker under both healthy and unhealthy conditions, taking into consideration the direct impact of COVID-19 on the respiratory system.

In the second experiment, investigated the fundamental frequency of the same speaker in healthy and unhealthy conditions. Given the common observation of throat inflammation during a COVID-19 infection, this study explored how such inflammation in the organs involved in fundamental frequency production could result in distinct measurements for the same speaker.

The third experiment delved into jitter value analysis, where jitter represents the variability or irregularity in the timing of vocal folds during phonation. Health compromises, such as respiratory infections or conditions like COVID-19, were considered as potential contributors to increased jitter values. Factors such as inflammation or swelling in the respiratory system, including the vocal folds, were examined for their impact on vibration patterns and the overall voice quality.

In the fourth experiment, the focus was on shimmer value assessment. Shimmer, indicating the cycle-to-cycle variation in amplitude during speech, was explored concerning compromised health. Conditions affecting the respiratory system, such as respiratory infections or lung diseases, were considered for their potential impact on airflow and subsequent variations in shimmer values. Vocal cord issues, including swelling, nodules, or paralysis, were also examined, as they can disrupt the regular vibration pattern of vocal cords and contribute to changes in shimmer values. Additionally, factors like muscle tension or weakness in the vocal tract, inflammation, and swelling in the vocal folds and surrounding tissues were studied for their influence on the coordination and control of vocal folds, further contributing to variations in shimmer values.

Since the identified features, including MFCC coefficients, fundamental frequency, jitter, and shimmer, exhibit variations with the speaker's health condition, they prove to be unreliable indicators. Consequently, it is imperative to explore alternative features beyond the mentioned ones for the effective training of a speaker recognition system.

The conceptual framework of this research is illustrated in Figure 3. A dataset containing voice recordings from five speakers was gathered in .wav format, featuring samples ranging from 3 to 10 seconds with a consistent sampling rate of 16 KHz. As the focus is on a text-independent system, diverse texts were utilized for both training and testing, enabling the model to learn speaker-specific attributes regardless of the spoken content. To address the small dataset size, a data augmentation approach was implemented, employing techniques like time stretching to introduce variations and expand the effective dataset size [20].

In this work, specifically, the time stretching technique was applied to augment the small dataset of speech signals. This technique involves altering the duration of the speech signal without affecting its pitch, introducing variations in temporal characteristics. By applying time stretching, new instances of the same speech content were generated with different durations, effectively enlarging the dataset and providing additional training examples for the speaker recognition system. No preprocessing was applied to the voice samples collected from the speakers. Following the modification of the dataset, the next step involved converting all voice samples into spectrograms. Convolutional autoencoders excel with image inputs, and representing voice samples as spectrograms allows for effective utilization of this architecture. Therefore, all voice samples were transformed into spectrograms after the database modification.

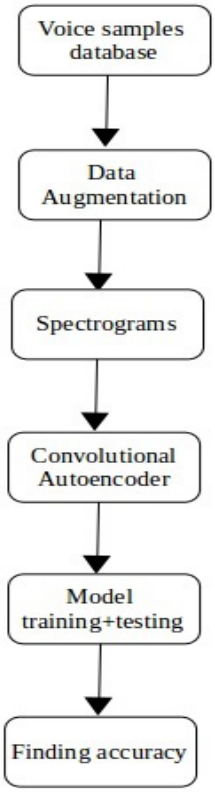

Figure 3. Proposed Framework

Convolutional Autoencoders (CAEs) harness the capabilities of convolutional operators to effectively capture spatial information. In contrast to conventional methods that involve manual engineering of convolutional filters, CAEs empower the model to autonomously learn optimal filters that minimize the reconstruction error. This learning ability positions CAEs at the forefront of unsupervised convolutional filter learning. In the realm of computer vision tasks, CAEs demonstrate proficiency in acquiring concise and meaningful representations of input data by leveraging the learned filters. These acquired features can then be employed for various tasks, including classification or any undertaking requiring a succinct representation of the input.

Although CAEs fall under the category of Convolutional Neural Networks (CNNs), a fundamental distinction sets them apart. CNNs are typically trained end-to-end, aiming to learn filters and amalgamate features for the classification of input data. On the contrary, CAEs specifically concentrate on learning filters tasked with extracting features used in the reconstruction of the input. This distinction underscores the unique purpose and objective of CAEs in comparison to traditional CNNs.

The merits of utilizing convolutional autoencoders are manifold. They exhibit proficiency in extracting high-level features from raw audio signals, contributing to more accurate speaker recognition compared to conventional feature extraction techniques. Furthermore, CAEs effectively filter out noise and other distortions from audio signals, enhancing the robustness of speaker recognition systems in noisy environments. Additionally, speaker recognition systems employing CAEs do not necessitate physical contact with the user, rendering them non-intrusive and convenient to use. Another noteworthy advantage is the adaptability of convolutional autoencoders to new data, allowing them to accommodate new speakers and dialects. This adaptability ensures that the system can continually enhance its accuracy over time [21]. Therefore, the utilization of convolutional autoencoders in this proposed methodology serves to address existing research gaps.

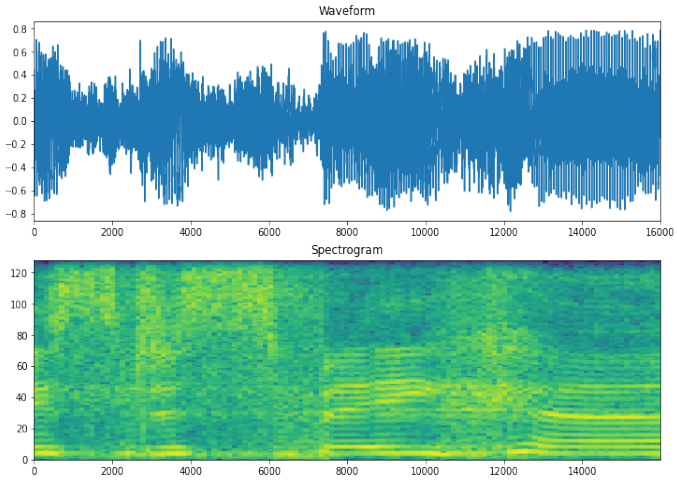

Figure 5. Voice Sample Visualization through Spectrogram Representation

As illustrated in the block diagram, all voice samples undergo conversion into spectrograms for training input. Figure 5 visually represents the voice sample transformed into a spectrogram. The activation function plays a pivotal role in determining system performance by introducing non-linearity into cells. It decides the involvement of neuron cells and holds significance in decision-making processes. While various activation functions exist, such as sigmoid, the proposed model adopts Rectified Linear Unit (ReLU) after experimentation, with mean squared error chosen as the loss function. The convolutional autoencoder employed in this experiment consists of two parts: the encoder (depicted in Figure 6) and the decoder (depicted in Figure 7). Pooling layers facilitate feature minimization, enhancing computational efficiency. Additionally, normalization is applied in conjunction with the activation function

Layer (type)	Output Shape	Param #
resizing_11 (Resizing)	(None, 32, 32, 1)	0
normalization_11 (Normaliza tion)	(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

Figure 6. CAE_Encoder

conv2d_transpose_23 (Conv2D Transpose)	(None, 28, 28, 128)	147584
conv2d_transpose_24 (Conv2D Transpose)	(None, 30, 30, 64)	73792
conv2d_transpose_25 (Conv2D Transpose)	(None, 32, 32, 32)	18464
conv2d_transpose_26 (Conv2D Transpose)	(None, 34, 34, 16)	4624
max_pooling2d_11 (MaxPoolin g2D)	(None, 17, 17, 16)	0
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

Figure 7. CAE_Decoder

4. Results

Throat infections, such as those experienced during COVID-19 or other respiratory illnesses, can induce alterations in pitch, Mel-frequency cepstral coefficients (MFCC), jitter, and shimmer features in human speech. These changes arise from several factors related to the impact of throat infections on the vocal apparatus. Inflammation caused by throat infections affects the vibratory characteristics of the vocal cords, thereby influencing the fundamental frequency or pitch of the voice. The inflammation also has implications for the resonant properties of the vocal tract, leading to variations in MFCC coefficients. Swelling of the vocal cords due to infections can disrupt their regular vibration, resulting in changes in pitch. The discomfort or pain associated with throat infections may prompt individuals to modify their vocal efforts, potentially causing alterations in pitch and MFCC coefficients. Additionally, congestion and increased mucus production linked to throat infections can impact vocal tract resonance and voice clarity, influencing pitch, MFCC coefficients, jitter, and shimmer.Throat infections can have broader effects on overall health, contributing to fatigue, weakness, and changes in respiratory function. These factors indirectly affect pitch and MFCC coefficients by influencing the coordination between the respiratory and vocal systems during speech production. It's crucial to note that the extent of these changes may vary based on the severity and type of throat infection, as well as individual differences in response to infections. Therefore, the analysis of pitch, MFCC features, jitter, and shimmer during throat infections provides valuable insights into the intricate effects of such infections on speech production and vocal health. However, based on the obtained results, these features appear to lack robustness for constructing an accurate and precise speaker recognition system. Consequently, further studies are warranted to identify more robust features within the speech signal.

The implementation of speaker recognition system based on CAE, utilizes the Python programming platform. A dataset comprising 50 voice samples from 5 distinct speakers has been collected. The dataset includes utterances ranging from 3 to 10 seconds, and diverse texts are employed for both training and testing. While the dataset encompasses a mix of languages, the primary focus is on developing a speaker recognition model based on the unique voice features of each speaker, making the language of utterances less impactful on the results. Each speaker exhibits distinctive characteristics, facilitating effective classification and achieving speaker recognition. To enhance the dataset's size for training and testing, an augmented dataset is employed. The entire dataset is partitioned into three sets: the training dataset, testing dataset, and validation dataset. Both training and testing experiments are conducted under two conditions - the matching condition and the mismatching condition. The matching condition occurs when the duration of training and testing utterances is the same, whereas the mismatching condition involves different durations. The accuracy curve for the matched condition is illustrated in Figure 8.

Throat infections, such as those observed in respiratory illnesses like COVID-19, can result in alterations to various aspects of human speech, including pitch, Mel-frequency cepstral coefficients (MFCC), jitter, and shimmer features. The underlying reasons for these changes are multifaceted. Inflammation induced by throat infections affects the vibratory characteristics of the vocal cords, leading to modifications in the fundamental frequency or pitch of the voice. Moreover, the inflammation can impact the resonant properties of the vocal tract, influencing MFCC coefficients. Vocal cord swelling caused by infections can disturb proper vibration, resulting in incomplete vocal cord closure or varying tension during vibration, ultimately causing pitch variations. Individuals experiencing discomfort or pain due

to throat infections may adjust their vocal effort by speaking with reduced intensity to alleviate discomfort, thereby affecting pitch and MFCC coefficients. Additionally, congestion and increased mucus production associated with throat infections can influence vocal tract resonance and voice clarity, leading to changes in pitches, MFCC coefficients, jitter, and shimmer.

Throat infections can have broader health implications, contributing to fatigue, weakness, and alterations in respiratory function, indirectly impacting pitch and MFCC coefficients by influencing the coordination of the respiratory and vocal systems during speech production. It's crucial to recognize that the extent of these changes may differ based on the severity and type of throat infection, as well as individual responses. Consequently, the analysis of pitch, MFCC features, jitter, and shimmer during throat infections offers valuable insights into the effects of such infections on speech production and vocal health. Despite these insights, further studies have revealed that these features are not sufficiently robust for constructing an accurate and precise speaker recognition system. Consequently, ongoing research is focused on identifying more resilient speech signal features.

Therefore, the spectrogram emerges as an alternative method for extracting features from speech signals. Unlike pitch, MFCC, jitter, and shimmer, the spectrogram provides a visual representation of the frequency content of a signal over time. By capturing the distribution of energy across different frequency bands, the spectrogram can offer a robust set of features that may prove more reliable for tasks such as speaker recognition. This approach leverages the time-frequency representation to characterize the unique patterns within speech signals, potentially overcoming some of the limitations associated with other feature extraction methods. As research continues, the spectrogram stands out as a promising avenue for enhancing the accuracy and precision of speaker recognition systems.

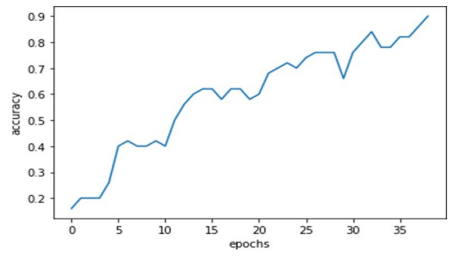

Figure 8. Accuracy Graph (Matched Condition)

Under matched conditions, the system has demonstrated a 92.4% accuracy rate. Figure 8 illustrates the accuracy curve in relation to the number of epochs, revealing that a smaller epoch rate is advisable for a smaller dataset. In Figure 9, the training loss and validation loss are depicted. It is essential for any system to achieve a perfect fit. Fortunately, there are no issues of overfitting or underfitting in this model. Figure 6 presents the curves for training loss and validation loss in the matched condition.

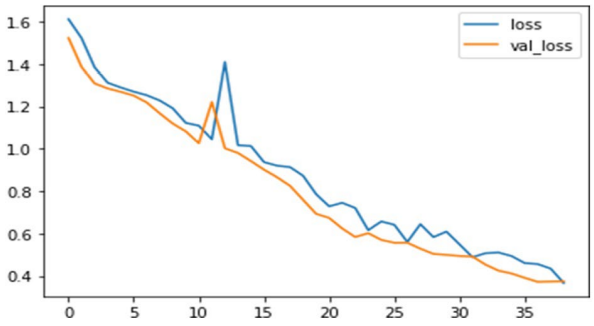

Figure 9. Loss Graph (Matched Condition)

Table 5. % Comparative Table of System Accuracy between Matched and Mismatched Conditions

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

The findings reveal that the system exhibits higher accuracy under matched conditions, where the test conditions align with the training conditions. Conversely, in the mismatched condition, characterized by significant differences between test and training conditions, the system's performance experiences a decline. Figure 10 provides a visual representation of the confusion matrix corresponding to the matched condition.

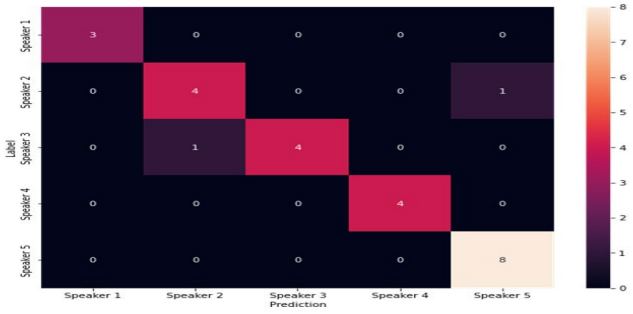

Figure 10. Confusion Matrix

Table 6 presents a comparative analysis between the proposed model and alternative methods. Figure 10 provides insights into the system's performance through a confusion matrix, illustrating the prediction rates for various labels, including Speaker1, Speaker2, Speaker3, Speaker4, and Speaker5. Under matched conditions, the system exhibits strong performance. However, in the case of mismatched conditions, the accuracy rate diminishes. Therefore, it is advisable to ensure uniform utterance lengths during both training and testing phases to enhance system performance.

Table 6. Comparative Analysis for the Proposed Model and Other Existing Approaches Using the 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

Examining pitch, MFCC feature jitter, and shimmer during throat infections offers valuable insights into the impact of such infections on speech production and vocal health. Despite this analysis, it has been determined from the results that these features lack robustness for constructing an accurate and precise speaker recognition system. As a result, further studies have been undertaken to identify a more resilient feature within the speech signal. In this research paper, a novel speaker recognition system is introduced, employing a convolutional autoencoder. The system demonstrated commendable success in matched conditions of utterances. However, its performance was found to be less satisfactory in mismatched conditions. Experimentation with various activation functions revealed that the ReLU activation function yielded superior results. Notably, the system utilized raw voice samples without preprocessing, showcasing resilience to background noise. A comparative analysis with established techniques like SVM and Random Forest, using the same dataset, showcased the proposed system's favourable accuracy rate, as indicated in Table 2. Evaluation metrics including Area under Curve, F1 score, CA, and precision were also compared. While prior studies in the related section predominantly employed MFCC features and clean datasets, this paper introduces novelty by representing speech signals as images through spectrogram conversion. Hence, the system used spectrograms as the feature for speech signals instead of MFCC. Additionally, a specifically collected dataset was utilized for this research. Experimental results identified a key challenge in the form of mismatched conditions between training and testing utterances. Thus, future research should address this issue to enhance the system's overall performance. Given that the system was tested with voices containing background noises, there is a potential avenue for improvement by exploring noise removal solutions. Furthermore, researchers are encouraged to explore the applicability of different types of autoencoders, such as denoising autoencoders and vanilla autoencoders.

References

[1] Mura, M. La., Lamberti, " Human-Machine Interaction Personalization: a Review on Gender and Emotion Recognition Through Speech Analysis." IEEE International Workshop on Metrology for Industry 4.0 & IoT, 319-323, (2020).

[2] Shelke, P. P., Wagh, K." Review on Aspect based Sentiment Analysis on Social Data". International Conference on Computing for Sustainable Global Development, 331-336, (2021).

[3] Ishak, Z., Rajendran, N., Al Sanjary, O. I., Mat Razali, N. "Secure Biometric Lock System for Files and Applications: A Review." IEEE International Colloquium on Signal Processing & Its Applications, 23-28, (2020).

[4] Soufiane H., Nikola N., Jamal, K, "Convolutional neural network vectors for speaker recognition." International Journal of Speech Technology, 24, 389–400, (2021).

[5] Tanu Singhal, "A Review of Coronavirus Disease-2019(COVID19)." Indian J Pedatr, 87(4): 281–286, (2020).

[6] Hu, H. R., Song, Y., Liu, Y., Dai, L. R., McLoughli, I., Liu, L," Domain Robust Deep Embedding Learning for Speaker Recognition." *IEEE International Conference on Acoustics, Speech and Signal Processing,* 7182-7186, (2022).

[7] Loina, L., "Speaker Identification Using Small Artificial Neural Network on Small Dataset." *International Conference on Smart Systems and Technologies,* 141-145, (2022).

[8] Lin, W., Mak, M. W., "Robust Speaker Verification Using Population-Based Data Augmentation." *IEEE International Conference on Acoustics, Speech and Signal Processing,* 7642-7646, (2022).

[9] Hasan, A., Abdulqader, S., Abdul Rahman Al-Haddad, S. Abdo, A. Abdulghani and S. Natarajan, "Hybrid Feature Extraction MFCC and Feature Selection CNN for Speaker Identification Using CNN: A Comparative Study." International Conference on Emerging Smart Technologies and Applications, 1-6, (2022).

[10] Ali, M. H., Jaber, M. M., Abd, S. K., Rehman, A., Awan, M. J., Vitkutė-Adžgauskienė, D., Damaševičius, R., & Bahaj, S. A, "Harris Hawks Sparse Auto-Encoder Networks for Automatic Speech Recognition System." Applied Sciences, 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 Journal on Audio, Speech, and Music Processing, 11, (2021).

[13] Bunrit, S., Inkian, T., Kerdprasop, N., Kerdprasop, K.," Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network." International Journal of Machine Learning and Computing, 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." Advances in Human-Computer Interaction, 1-11, (2023).

[15] Rashid Jahangir, Ying Wah Teh, Henry Friday Nweke, Ghulam Mujtaba, Mohammed Ali Al-Garadi, Ihsan Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. "Expert Systems with Applications, 171, (2021).

[16] Piotr Staroniewicz, "Influence of Natural Voice Disguise Techniques on Automatic Speaker Recognition",.Joint Conference - Acoustics,IEEE,(2018).

[17]] Douglas A.Reynolds and Richard C. Rose, "Robust text in- dependent speaker identification using Gaussian mixture speaker models." IEEE, (1995).

[18] Lap-Ching Keung, Kelly Richardson, Deborah Sharp Matheron, Vincent Martel-Sauvageau, " A Comparison of Healthy and Disordered Voices Using Multi-Dimensional Voice Program", Praat, and TF32,Journal of Voice,ISSN 0892-1997, (2022).
[19] Hengling Zhao, Yangyang Jiang, Shenghan Wang, Fei He, Fangzhou Ren, Zhong- hao Zhang, Xue Yang, Ce Zhu, Jirong Yue, Ying Li, Yipeng Liu, "Dysphagia diagnosis system with integrated speech analysis from throat vibration." Expert Systems with Applications, Volume 204, 117496, ISSN 0957-4174(2022).
[20] R. Jagiasi, S. Ghosalkar, P. Kulal and A. Bharambe." CNN based speaker recognition in language and text-independent small-scale system." International conference on IoT in Social, Mobile, Analytics and Cloud, 176-179, (2019).
[21] Tirumala, S.S., & Shahamiri, S.R., "A Deep Autoencoder approach for Speaker Identification." International Conference on Signal Processing Systems*, (2017).*