Khazaleh and Khrais
Journal of Engineering and Applied Science (2024) 71:13
https://doi.org/10.1186/s44147-023-00351-0

Journal of Engineering and Applied Science

RESEARCH Open Access

An investigation into the reliability of speaker recognition schemes: analysing the impact of environmental factors utilising deep learning techniques

Omar Ratib Khazaleh[1*] and Leen Ahmed Khrais[1]

*Correspondence:
omar.ratib12@gmail.com

[1] Yarmouk University, Irbid, Jordan

Abstract

This paper studies the performance and reliability of deep learning-based speaker recognition schemes under various recording situations and background noise presence. The study uses the Speaker Recognition Dataset offered in the Kaggle website, involving audio recordings from different speakers, and four scenarios with various combinations of speakers. In the first scenario, the scheme achieves discriminating capability and high accuracy in identifying speakers without taking into account outside noise, having roughly one area under the ROC curve. Nevertheless, in the second scenario, with background noise added to the recording, accuracy decreases, and misclassifications increase. However, the scheme still reveals good discriminating power, with ROC areas ranging from 0.77 to 1.

Keywords: Speech recognition, Deep learning, Convolutional neural networks, Recording conditions, Background noise, Speaker recognition dataset, ROC curve, Real-world scenarios

Introduction

Automatically ambient sound classification has recently gained attention as a rapidly developing topic with various uses. In this scope, a lot of research in related fields, including speaker, recognition of voices, and music, have been performed, but an obvious oversight is remarked in the study of how to classify natural sounds. The question of whether these techniques are appropriate in other areas, including sound classification, arises in light of developments within image classification, notably the application of convolutional neural networks enabling precise besides scalable image recognition.

Spectrograms, a plot which displays a sound's frequency range as well as fluctuations over time, have been shown to remain an effective tool for examining sound properties quickly [1].

The study of speaker signals has received significant scientific attention, intending to resolve the term ASR which stands for the automated speaker recognition problem [2]. Nevertheless, the primary area of this study, sound event recognition, has increased

interest in research and benefited from the extensive ASR research carried out over the years.

Sound occurrences are text-labelled audio clips correlating to any interesting action or occurrence. Various audio sources, such as people, animals, or certain environmental sounds can cause these occurrences. By broadening the behaviour of linked sound events, signal processing and machine learning methods are used to detect these occurrences in the numerical world. The ability to classify newly witnessed events and recognise sound events from previously undetected acoustic source signals is a significant benefit of this generalisation. Sound event identification is being studied in many application fields, including automated acoustic tagging, acoustic surveillance, medical monitoring, and determining the auditory context of surroundings. Additionally, several firms that serve the requirements of customers and other enterprises have included sound event recognition as a component of the associated business models [3].

With the use of gradient descent and optimisation approaches, DL, which is a term that stands for deep learning, focuses on identifying the most suitable variables of both linear and non-linear functions. Since DL approaches often involve training based on the appropriate datasets, the total approach is the supervised learning procedure. However, new strategies for changing the whole forecasting procedure to stand unsupervised have been discovered by researchers [4].

Due to their better accuracy and systematic feature engineering, DL approaches have experienced a comeback. However, the theoretical backlog and unexplained DL models are the present limitations of DL techniques. Due to the associated superior prediction skills, businesses with huge datasets firmly rely on DL approaches despite these drawbacks [5].

Investigators take advantage of the speech recognition domain's perception of the durability of DL frameworks. DSR, which stands for Deep speaker recognition, is a term used to describe the speaker recognition procedure used in conjunction with DL methods [6].

DL is a modest technique exploring its competencies in several domains, which contains speaker recognition. Consequently, a fast-paced development is observed, and new thoughts are anticipated. Though DSR presents the new speaker recognition area platform, previous methods will be outdated soon. Additionally, the DL and ML theoretical grounding is founded on speaker recognition instead divergent. Therefore, DL-based experiments are required to adequately illustrate the current state of the DL built on speaker recognition. The use of Gaussian mixture models (GMM), Hidden Markov models (HMM), and Universal Background Models (UBM) has helped speaker recognition techniques advance. The GMM-UBM scheme combination results in the development of the i-vector, which is used unaltered as the speaker recognition starting point across several platforms. However, deep learning techniques are state-of-the-art in some sectors, including the processing of languages, computer vision, communication, networking, and defect detection. Consequently, researchers are presently concentrating on speaker recognition schemes utilising deep learning [7, 8].

Speaker recognition schemes are widely utilised in critical implementations, but several factors can endanger their dependability. One important factor is the equipment recording utilising to obtain the speech information. Different recording setups, audio devices, and

microphones can present differences in the signal of recorded speaker voices, impacting the robustness and accuracy of speaker recognition schemes. Moreover, environmental factors, for instance, room acoustics, reverberation, and background noise, can further damage these schemes' performance [9, 10].

Problem statement

Understanding the dependability of speaker recognition schemes under several recording devices and environmental situations is crucial for enhancing their accuracy and usability. Current research has mainly focused on advancements in algorithmic, while the recording device's influence and environmental conditions have received restricted attention. Later, there is the requirement for an in-depth investigation to evaluate the dependability of these schemes and identify possible challenges and explanations in real-world situations.

The paper will tackle this issue by examining how environmental factors and recording equipment affect deep learning founded on speaker recognition schemes. By evaluating the resilience and performance of these schemes across several environmental context conditions, the study intends to offer practical suggestions and valuable insights for improving the speaker recognition scheme's precision and dependability in real-world utilise cases.

This study focuses on implementing deep learning techniques employing MATLAB and Python software alongside convolutional neural networks (CNNs) to examine the reliability of speaker identification systems because of recording equipment and ambient conditions. The study aims to assess the effectiveness and reliability of speaker recognition systems built on deep learning under various ambient situations.

The paper uses a methodical approach to provide its material in a clear and useful way. Beginning with the introductory section, which sets the scene by highlighting the significance of speaker recognition methods and the emerging subject of automatic environmental speaker categorisation, the rest of the chapter introduces the key concepts.

Progressing, part 2 demonstrates the problem statement, concentrating on the idea of sound measures and their recognition in the digital domain. Part 3 delivers a related work overview, highlighting the previous research significance in the area and demonstrating the relevance of speaker recognition schemes and the rising interest in the classification of automatic environmental speaker voice. Part 4 summarises the methodology working for sound happening detection, illumination, the processing of signals and algorithms of machine learning used.

In part 5, the simulation design is comprehensive, offering an understanding of how the speaker event detection scheme performance was assessed.

The following part, part 6, shows the results and a discussion, examining the simulation outcomes and contrasting them with previous studies to attract expressive conclusions. Lastly, part 7 provides an exhaustive conclusion outlining the key results, discussing their possible implications, and signifying paths for upcoming research.

Literature review

Systems for identifying speech or speaker voices are often employed in a wide variety of applications, including biometric identification, virtual assistants, and security systems. Nevertheless, recording and speaker-identifying technology and environmental conditions can impact these systems' re [11].

Voice signals constitute a universal type of interaction that always convey valuable information, including the accent, gender, mood, and other distinctive traits of the speaker. When phone conversations are made, even if the speakers are not there physically, researchers can distinguish between speakers according to these distinctive traits, referred to as voice biometrics. Through these traits, robots can develop a human-like understanding of speaker utterances. Speaker identification (also referred to as SI) is the technique of employing a machine to identify a speaker based on the acoustic features of a particular speech [12].

Speaker recognition seeks to determine the speaker based on traits such the accent, speaking style, and pitch, as opposed to speech recognition, which focuses on turning audio into text. These technologies have a variety of uses, including aiding people with impairments and enhancing human–computer interactions. This article examines the components of speaker identification systems, such as feature extraction, preprocessing, and speech modelling.

The diversity of speech signals and the impact of recording equipment and transmission techniques are also discussed, as are the difficulties in recognising speakers as a result. In order to close the knowledge gap and increase the relevance of these technologies, the related work part covers earlier surveys and literature reviews that investigated feature extraction techniques, deep learning architectures, and various other speaker recognition-related topics.

The two most popular applications in speech processing that researchers employ for analysing communication are speaker recognition and speech recognition [13]. It is crucial to comprehend the distinction between speaker and speech recognition before getting deeper into the fundamentals of speaker recognition. While speaker or voice recognition focuses on the speaker instead of the words being uttered, speech recognition stays involved regarding the words being uttered.

Speaker recognition vs speech recognition

People with a range of disabilities, such as those with physical limitations who find it challenging, uncomfortable, or impossible to speak, can benefit from speech recognition or impossible to type, as well as people with dyslexia who have trouble reading and spelling words [14].

Speech recognition is concerned with turning audio into text; therefore, the language and text corpus significantly impact how well it works. On the contrary, speaker identification aims to identify the speaker. Some characteristics which add to the variances include accent, speaking style, and pitch [15]. Application areas for speaker recognition technologies include biometrics, assurances, and interaction between humans and computers [16].

According to recognition, objective, focus, and application, Table 1 compares and contrasts speech recognition and speaker recognition [16]. The relevance of speech recognition technologies has grown due to advancements in several industries, particularly when determining a person's identification.

Table 1 A comparative analysis of speech and speaker recognition [16]

Features	Speaker Recognition	Speech Recognition
Recognition	Recognises who is speaking by measuring voice pattern, speaking style, and other verbal traits	Recognises what is being said and converts them into text.
Purpose	To identify the speaker	To identify and digitally record what the speaker is saying.
Focus	Biometric aspects of the speaker, such as pitch, intensity, etc., to recognise him/her	Vocabulary of what is being said by the speaker and turns the words into digital texts.
Application	Voice biometrics	Speech to text.

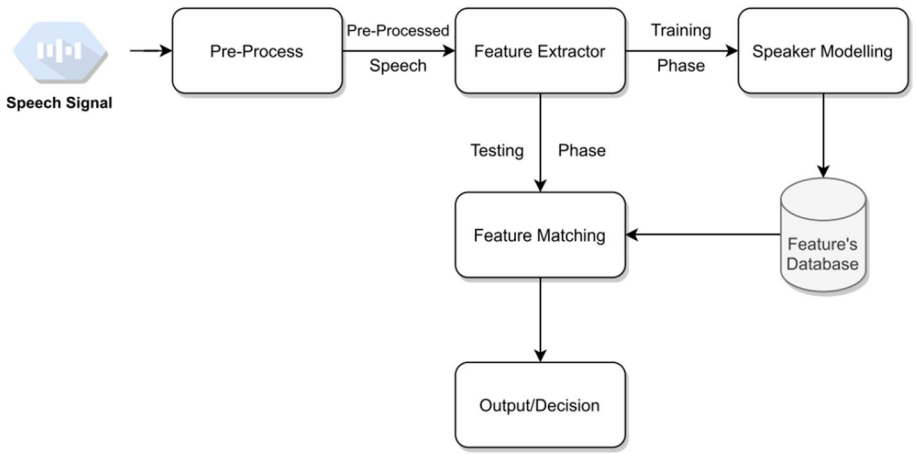

Fig. 1 The fundamental design of speech recognition systems

Structure of speaker recognition

Training and testing are the two steps of deep learning architectures' training. In speech identification techniques, registration is frequently used to describe the training phase. Such unique processes of the speaker identification platform are discussed in this section. Speech modelling, feature extraction, and preprocessing are the three basic facets of a speaker identification system. Figure 1 depicts a basic diagram of voice recognition systems.

- Preprocessing: The first step in an automated speech identification model is preprocessing. To create an efficient and dynamic ASR system, executing this operation on the speech signal input is essential. This part of the speaker identification system is where the speech signal is initially cleaned. The remaining non-speech components are then removed, cleaning up the signal. Endpoint identification and pre-emphasis are the next preliminary tasks to be finished [16].

Three different preprocessing methods for speech data have been identified: spectrogram, mel-filterbank, and MFCC. The procedure for obtaining the features is shown in Fig. 2 [4]. The use of DL in speaker recognition systems is widespread. As datasets and DL techniques develop, researchers are learning many new features of

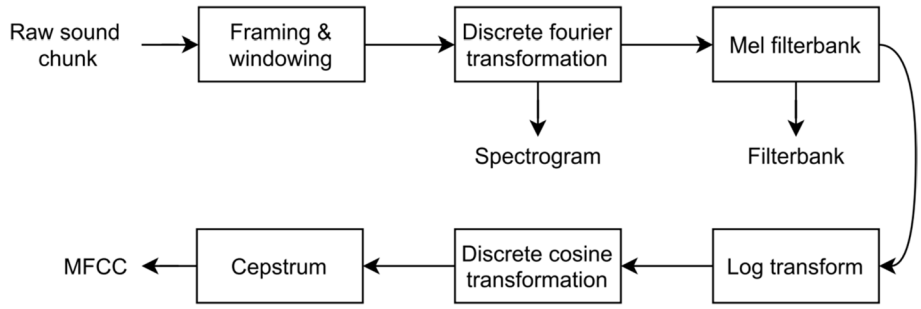

Fig. 2 The diagram shows the procedures needed to extract a spectrogram [4]

speech recognition systems. As a result, several methods have been investigated in speaker identification systems.

- Front-end preprocessing, often referred to as feature extraction, is used throughout speaker recognition systems' training and testing phases. It uses feature vector or numerical descriptor sets to transform digital voice signals. The key elements of the speaker's speech are represented in these feature vectors [4].

According to Nolan [17], a feature parameter should ideally [9]: (1) demonstrate minimal within-speaker variability with significant between-speaker variability, (2) refuse any attempts at disguising or mimicking, (3) are prevalent often in pertinent sources, (4) have strong transmission, (5) be rather straightforward for collecting and measuring. Despite being stated in the context of forensic speaker identification, these characteristics are generally applicable.

- Speaker modelling: Speaker recognition algorithms are created using modelling techniques to match speaker speech features. Speaker models are characterised as processes that combine increased speaker-specific information with reduced volume. State speaker models are created throughout training or enrolment by repeating the specific traits of a contemporary speaker. In the recognition state, when used for tasks like recognition or verification, the speaker model contrasts with modern speaker architecture [4].

Challenges in speaker recognition

Speech, unlike other biometrics such as fingerprints, facial features, irises, gait, as well as hand geometry, is a performance-based biometric. The speaker's identity is primarily embedded in the manner of speech rather than the content spoken. This characteristic introduces a high level of variability in speech signals.

Even the same person does not always talk in the same way; this is referred to as intraspeaker variability or style-shifting. Additionally, transmission methods and recording devices further contribute to the challenge. For instance, voice recognition through a phone or when a person has a cold, performs another task, or speaks with different vocal efforts like whispering or shouting can be difficult [9].

Related work

In the recent systematic literature review [18], authors focused on identifying significant feature extraction methods within the past 6 years. The review aimed to provide recommendations founded on the investigation and address three important questions in the domain of speaker recognition. The authors explored the basics for optimum features through feature extraction, deriving methods and architectures that have been traditionally successful.

Additionally, they discussed several challenges faced in the field. In another study [19], the authors reviewed various deep-learning approaches for speaker identification. They described the deep learning applications' structures and algorithms, which have attained modern accuracy. The goal was to link and enhance the deep learning significance and the knowledge gap in speaker recognition research.

A complete literature survey conducted by the authors in [20] delved into ASR systems, categorising the speaker recognition modules and presenting various models for each component. They also provided a concise overview of the vast implementations of SR Systems and expounded on the associated issues and challenges. Authors in [21] brought forth automatic speech recognition (ASR) techniques, highlighting the notable emphasis on text-independent recognition in speech and speaker recognition between the 1980s and 2009, with a particular focus on more recent methods introduced about 2009.

This research aided as an analytical survey of the domain of speaker recognition, elucidating key inquiries and explanations. In [22], authors evaluated significant subdomains within speaker recognition, including speaker diarization, verification, and identification, emphasising deep-learning-based methods.

The paper extensively explains modern deep learning-based techniques for feature extraction and ASR algorithms. Moreover, other surveys in the speaker recognition domain have been introduced at various points in time [23]. Given the limited coverage of those surveys, a wide study was necessary to consolidate knowledge in this domain.

Methods

This chapter illustrates the procedure which was used to apply the classification system under various conditions, including data source identification, dataset description and the conducted method:

Data source

A full collection of audio recordings used in this study from diverse speakers can be found in the Speaker Recognition Dataset, which is accessible at the Kaggle online website source [24]. Researchers and developers working on speaker analysis and identification might benefit greatly from the resources provided by this dataset, which was created expressly for tasks involving speaker recognition.

Data description

The Speaker Recognition Dataset which includes speaks from numerous speakers, is described in the following breakdown of the information offered:

1. Folder names: The dataset includes speeches from four distinct speakers. Each speaker gets their folder with their name on it. The provided dataset has undergone filtration, resulting in the segregation of data into three distinct categories characterized by individuals of similar gender and tone of voice. This categorisation enables a systematic examination and evaluation of the program's proficiency in classifying and identifying analogous voices, as well as its capability to discern the presence of noise within the dataset.
2. The audio files break the remarks into digestible one-second audio segments. Each audio file contains a sampling rate of 16,000 samples/s and is PCM (ulse-code modulation) encoded. Accordingly, the audio records 16,000 sound samples each second, precisely depicting the speech.
3. Speech segments: Each speaker's speech may be recreated by combining the divided audio files from 0 to 1500. wav. According to this, the lectures were likely first recorded as one large audio file that was then divided into smaller chunks for simpler administration and analysis.
4. Background noise: The information set also contains a subdirectory called 'background_noise'. These audio files are not lectures but sounds that may be heard in the speaker's setting or nearby. The audience's cheers and laughing might be among these background noises. These files are included to mimic actual scenarios when various background noises support speaker voices.
5. Training data: The background noise audio files and the chunked speaker recording files can be mixed during training. Combining these two techniques enables the speaker identification system to become proficient at identifying and differentiating speaker voice from various speakers over realistic background disturbances.

This dataset offers a selection of talks by well-known speakers, divided into one-second auditory chunks, with the capability to insert background sounds for purposes of training in order to imitate real-world situations. In summary, the deliberate division of the dataset into discrete voice classifications enhances the breadth and precision of the assessment procedure, providing a thorough grasp of the program's advantages and shortcomings for voice classification and noise reduction.

Recording properties

Based on the provided information on the Kaggle website about the data context, the following are additional details about the recording dataset characteristics:

1. Audio duration: The dataset's audio files are all exactly one second long, making segmentation and analysis consistent.
2. Sampling rate: The audio recordings are sampled at a constant rate of sixteen thousand samples per second. The number of samples (data points) that were captured

to record the audio every second is indicated by the sampling rate. Greater accuracy and quality of audio are sometimes indicated by greater sample rates.

3. Encoding: PCM (pulse-code modulation) is used to encode the audio files. A popular format that guarantees an accurate reproduction of the audio signal is PCM. It is renowned for being of the highest calibre and being lossless, which makes it appropriate for jobs involving speaker recognition.

Important details regarding the Encoding, calibre, and structure of the audio data in the dataset are provided by these recording attributes. With this knowledge, researchers and developers will be able to comprehend the features of the recordings and successfully utilise them for speaker analysis and recognition.

Applied method

This paper's methodology emphasises analysing the speaker recognition schemes' reliability in regards the subsequent phases (Fig. 3):

Data collection

The first phase of the work is to gather the data for speaker audio. A detailed directory path involving the speaker audio files is set. Moreover, a directory is formed to store the spectrogram images. The work methodology involves finding a list of WAV files in the definite directory to process the audio data.

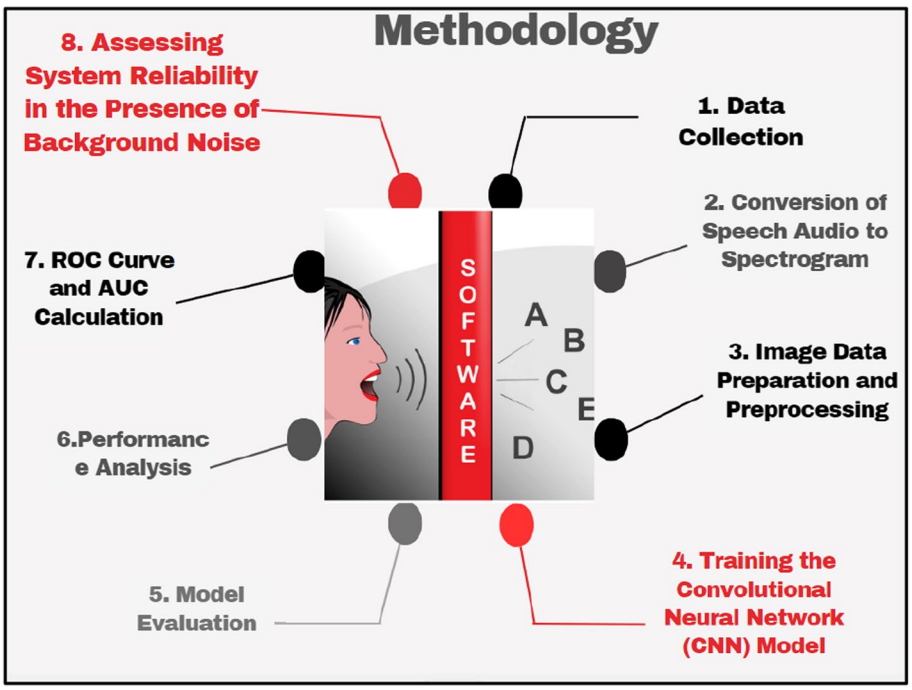

Fig. 3 Methodology phases

Speaker audio conversion to spectrogram

To investigate the speaker audio data, it is transformed into spectrogram images. Spectrograms offer a visual representation of the frequency gratified of the audio over the period. For example, the overlap ratio and window length are specified for the spectrogram calculation. These parameters establish the frequency and temporal resolution trade-off in the finding spectrograms. Every audio file in the specified directory is processed iteratively. The audio is read using the 'audioread' function, and the spectrogram is calculated using the 'spectrogram' function.

The finding magnitude spectrogram is transformed to the dB scale utilising logarithmic conversion. Then, the spectrogram is plotted, with frequency on the y-axis and time on the x-axis. The spectrogram is stored as a PNG image in the before created image directory. This conversion from speaker audio to spectrogram allows additional analysis and processing of the speech data utilising image-based methods.

Figure 4 demonstrates a sample of a speaker spectrogram transformed in MATLAB.

Image data preprocessing and preparation

After finding the spectrogram images, the next phase is to make the image dataset for training. The spectrogram images are resized to an unchanging size of 50×50 pixels utilising the OpenCV ('cv2') library. Resizing the images confirms that they have matching dimensions, essential for training CNN. The grayscale images are transformed to grayscale utilising the 'cv2.imread' function with the 'cv2.IMREAD_GRAYSCALE' flag.

This extends the image representation by eliminating colour information while retentive the essential frequency content. The grayscale and resized images and their

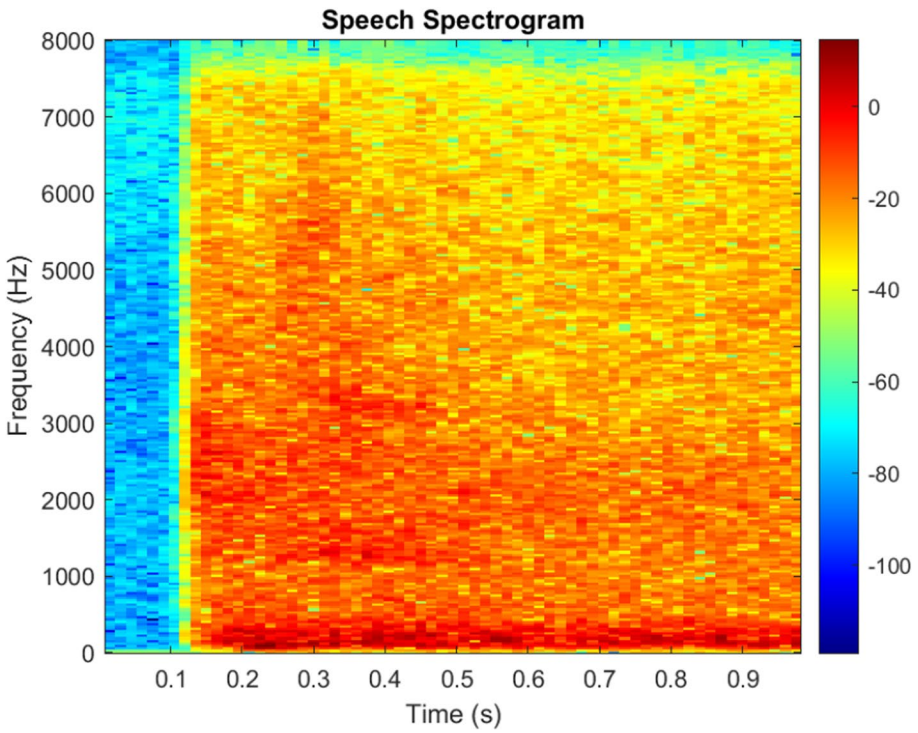

Fig. 4 Sample of speaker spectrogram

consistent class labels are saved in a list named 'training_data'. Random shuffling is practical to the training data to confirm that the model shows varied samples during training.

Before training the CNN model, the image data requirements are to be preprocessed. The feature labels ('y') and vectors ('X') are detached from the 'training_data' list. The feature vectors, showing the spectrogram images, are redesigned into a 4D array utilising NumPy ('np. array'). This conversion ensures that the input data is well-matched with the input shape predictable by the CNN model. Furthermore, the image's pixel values are scaled to a normalised range of [0, 1].

This scaling is completed by dividing the pixel by 255, which regulates the pixel strengths across all facilitates and images convergence during training.

Training the CNN model

The ready image dataset is utilised to train a CNN model. The CNN architecture is built utilising TensorFlow's Keras Sequential API (Application Programming Interface). The model includes the convolutional, max pooling, and fully connected layers. These layers allow the model to learn hierarchical illustrations and capture spatial features in the spectrogram images. The model is accumulated, which is operative for training deep neural networks.

The task involves speaker recognition. Hence the function for binary cross-entropy loss is used as the benchmark for optimisation. The precision metric is selected to evaluate the performance model during training. The system is trained for specific epochs using the training data (10 in this work). The validation data, consisting of the training data, is utilised to monitor the performance model and avoid overfitting.

Model evaluation

When the CNN model is trained, it is estimated using different techniques. Initially, the trained model is stored for future utilise. Formerly, a function named 'preprocess_image' was specified to preprocess a single image for testing. This function achieves the essential steps, for instance, reading the image file, resizing it, and normalising the pixel values, to make it for input to the trained model.

The user is prompted to choose an image file for testing, and the chosen image is preprocessed utilising the 'preprocess_image' function. The preprocessed image is provided in the trained model for the forecast. The model forecasts the class probabilities for the specified image, demonstrating the likelihood of individual classes. The forecast class label is decided and found on the predicted probability. The predicted label is outputted as the output, demonstrating the recognised speaker in the image.

Performance analysis

A broader performance analysis is conducted after evaluating the model on each test sample. Predictions are complete on the test set, which contains an image dataset portion that was not utilised for training. The predicted classes are flattened and rounded to get binary predictions. A confusion matrix is created utilising the 'confusion_matrix' function from the 'sklearn.metrics' module. The confusion matrix offers insights into the classification performance model, and display the true positive number, true negative, false negative, and false positive predictions. The confusion matrix is visualised

utilising a heatmap plot with explanations, created utilising the 'seaborn' and 'matplotlib' plot libraries. This visualisation aids in understanding the distribution of the prediction across various identifying classes for any possible misclassifications.

ROC curve and AUC calculation

Creating a ROC curve allows for further evaluation of the performance model. The true positive rate (TPR) and false positive rate (FPR) are calculated using the probability that the model predicts for various classification thresholds. Using the 'roc_curve' function from the 'sklear'. The metrics module, the FPR, TPR, and related thresholds are obtained.

The area under the ROC curve (AUC) is intended to utilise the 'auc' function from the same module. The TPR is on the y-axis, and the FPR is on the x-axis. The curve represents the trade-off between FPR and the real positive rate at different threshold values. The AUC value calculates the overall model performance, with a higher value demonstrating better discrimination aptitude. The AUC and ROC curve deliver valuable insights into the classification performance mode and can be utilised to compare threshold settings and various models.

Assessing scheme reliability in the background noise presence

To improve the scheme evaluation reliability in speaker recognition, this research methodology includes adding background noise and evaluating scheme performance under these challenging situations. The methodology includes collecting a separate dataset of different background noise samples generally encountered in real-world situations.

These noises are then joined with the original speaker audio dataset to generate augmented training and testing datasets. The augmented speaker audio, including the covered background noise, is transformed into spectrogram images. The image data is prepared and preprocessed as before, counting resizing, grayscale conversion, and pixel value normalisation. The CNN model is trained, utilising the augmented and preprocessed image dataset. Throughout the evaluation, the trained model is tested on the augmented testing dataset, which covers speaker voice samples with overlaid background noise at various signal-to-noise ratios (SNRs). The system performance is assessed by measuring precision. This comprehensive analysis allows a robust evaluation of the speaker recognition scheme's reliability under challenging and realistic acoustic environments.

Results and discussion

The outcomes of the two scenarios were examined to assess the designed system, in terms of four cases. The first scenario was investigated without considering the effects of outside noise on the sounds.

In the four cases, three of them have two sounds, and the fourth one has three sounds in each scenario, which were evaluated to see how well the DL technique could identify each sound. For each scenario, a ROC plot and confusion matrix were produced.

Scansion one: without noise

The confusion matrix and ROC of speaker sound classification are displayed in Figs. 5 and 6, respectively: 'Speaker one', 'Speaker two'.

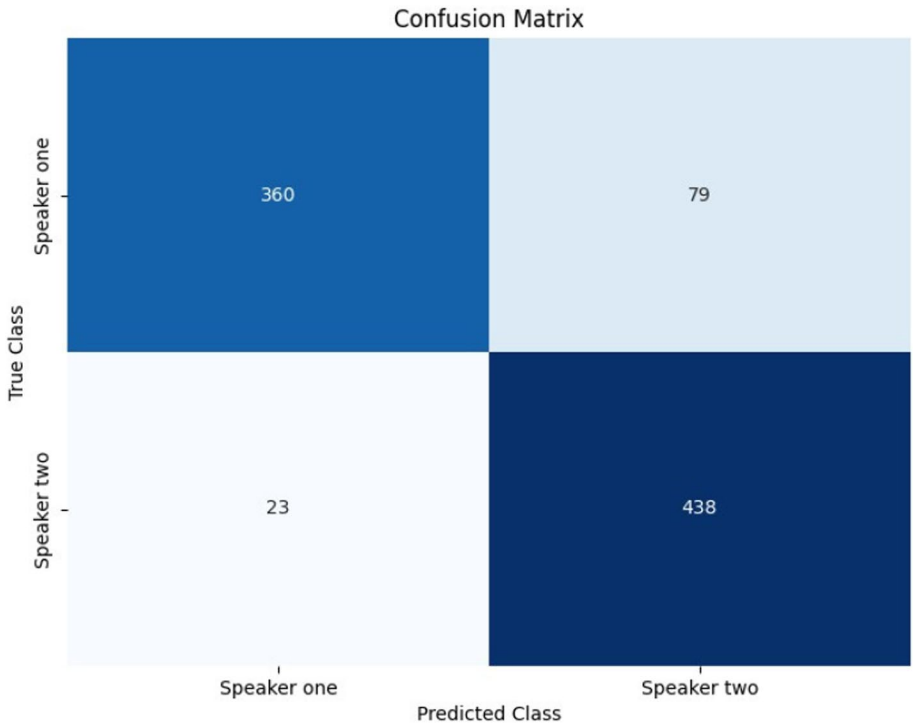

Fig. 5 Confusion matrix for case one

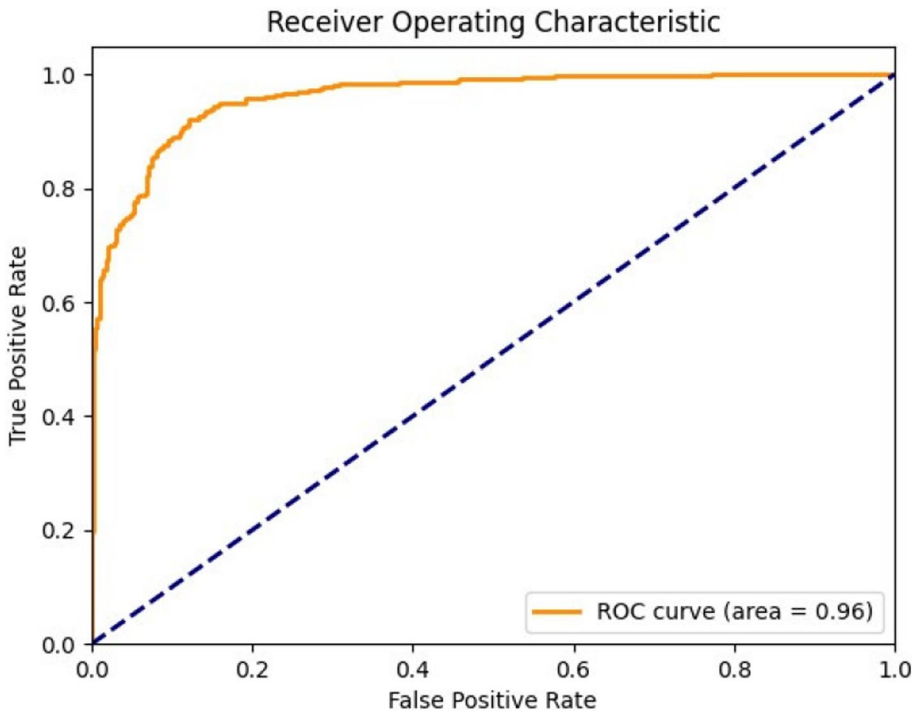

Fig. 6 Receiver operating characteristic for case one

Three hundred sixty instances in the Speaker 1 category were found to be correctly recognised, whereas 79 cases were found to be incorrectly classified. Of the Speaker two category, 438 instances were correctly recognised, while 23 cases were incorrectly categorized. The confusion matrix displays an examination of the accuracy of the system for every category.

The ROC plot, which showed an area of 0.96, demonstrated the system's high level of accuracy and discriminating power. The ROC curve presents the system's performance at different categorisation criteria visually. These results indicate that both groups were correctly identified without external noise and with a high degree of accuracy. The system's ability to discriminate well is indicated by its high AUC value, which shows that it can effectively distinguish between them.

The confusion matrix and ROC of speaker sound classification are displayed in Figs. 7 and 8, respectively: 'Speaker one', 'Speaker three'.

Results from the figures above were examined concerning the created system. The categories 'Speaker one' and 'Speaker three' were taken into consideration. The following conclusions are drawn from the confusion matrix:

Four of the 472 instances in the Speaker one category were incorrectly categorized. In the same way, 8 of the 408 instances in the Speaker three category were incorrectly categorized. The confusion matrix demonstrates that the algorithm achieved a 99.15% classification accuracy for both categories of events.

On the ROC plot, one area was also discernible. This demonstrates that the system differentiated between the two groups with perfect discriminating ability and optimal performance. These results indicate that, in the absence of external noise, the

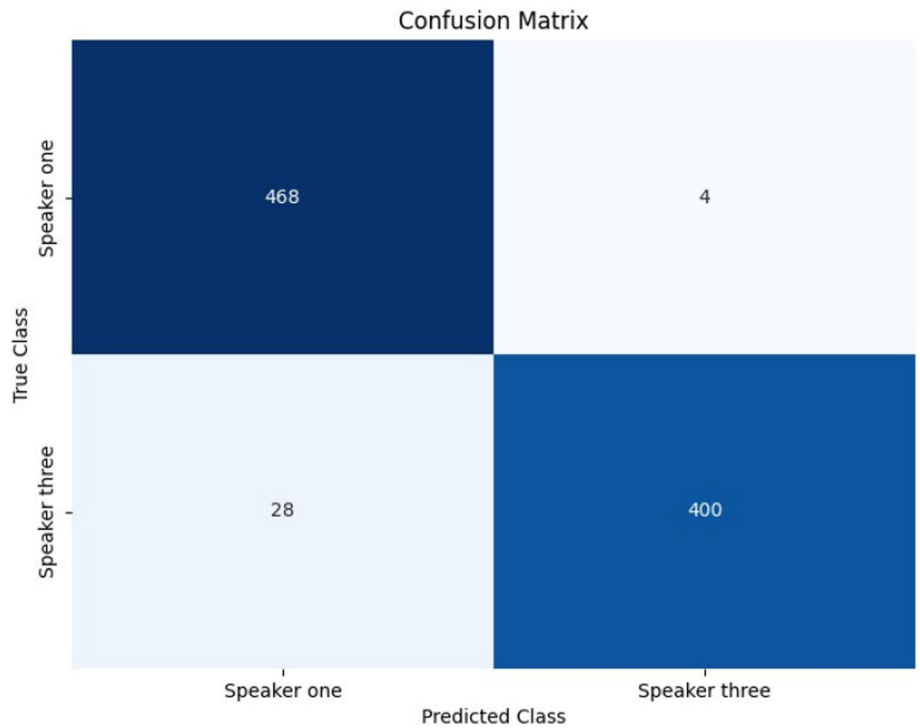

Fig. 7 Confusion matrix for case two

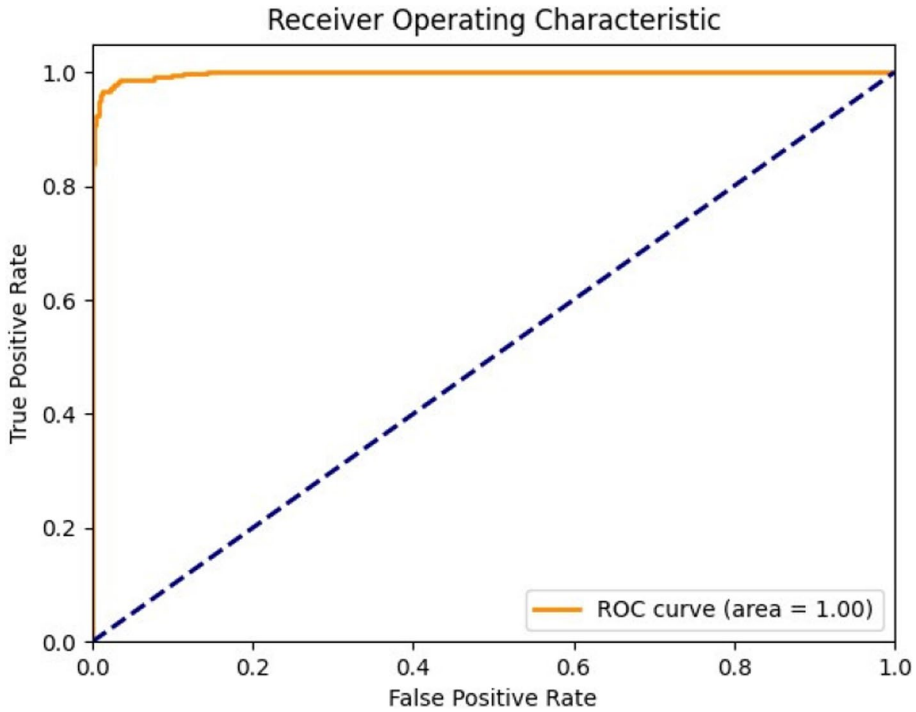

Fig. 8 Receiver operating characteristic for case two

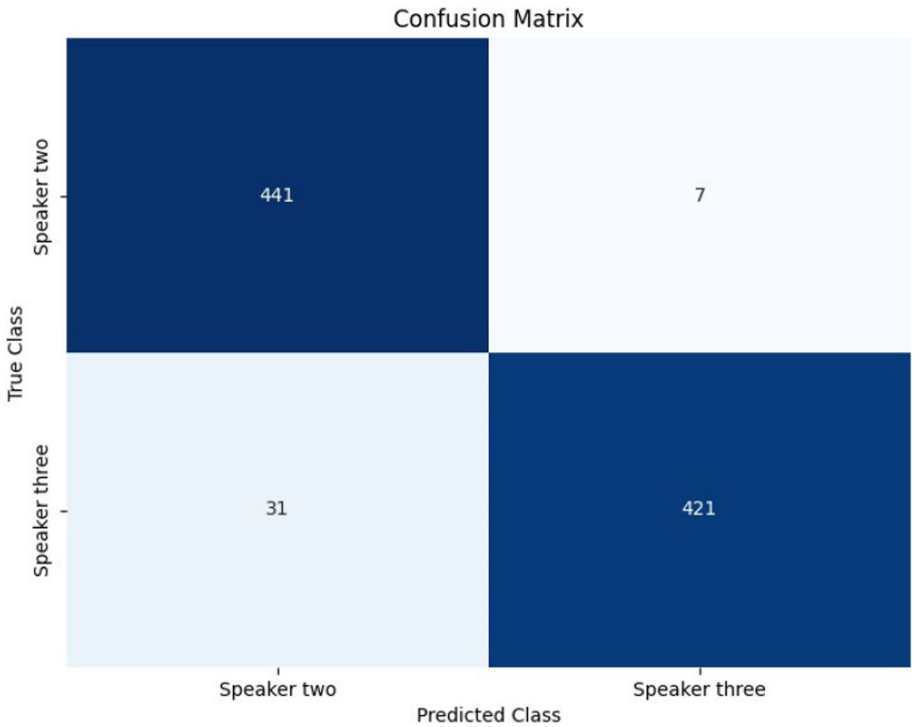

Fig. 9 Confusion matrix for case three

suggested approach fared exceptionally well for the Speaker one and Speaker three categories. The system can dependably and efficiently detect speakers' voices in a variety of ambient situations, as demonstrated by its ROC area of 1 and absolute classification accuracy.

The confusion matrix and ROC of speaker sound classification are displayed in Figs. 9 and 10, respectively: 'Speaker three' and 'Speaker two'.

'Speaker three' and 'Speaker two' were the categories taken into consideration in the modified example. There were no occurrences of the 441 cases in the Speaker 2 group that were incorrectly categorized. In a similar vein, 31 of the 421 cases in the Speaker 3 group were incorrectly categorized. The system functioned best when it was able to distinguish between the two groups, as indicated by the ROC plot's area of 1. In Fig. 11, the confusion matrix for speaker sound classification is displayed: ['Speaker three', 'Speaker two', 'Speaker one']:

With regard to the new case including the categories 'Speaker three', 'Speaker two', and 'Speaker one', 351 of the 448 cases were correctly recognised, but 97 of the cases were wrongly classified. In the Speaker two category, out of 455 cases, 440 were properly recognised and 15 were incorrectly categorized. 390 cases out of 447 in the Speaker three group were accurately recognised, while 57 cases were mistakenly classified. The confusion matrix illustrates how well the algorithm classified speakers' sounds for each category.

Scenario two: with background noise

The confusion matrix and ROC of categorising speaker sounds in the presence of background noise are displayed in Figs. 12 and 13, respectively: ['Speaker one', 'Speaker two', and 'noise'].

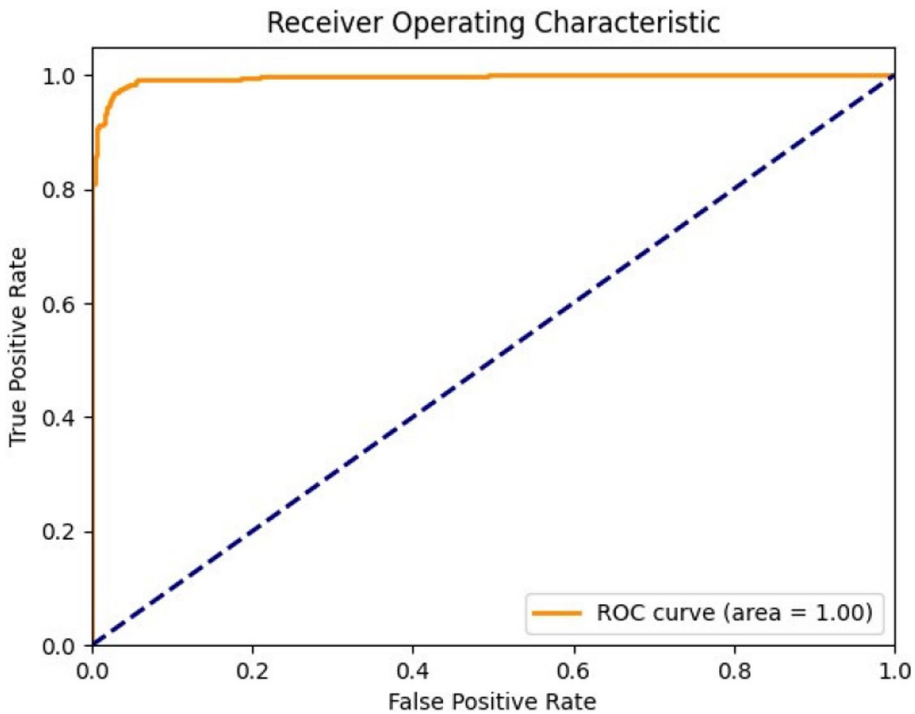

Fig. 10 Receiver operating characteristic for case three

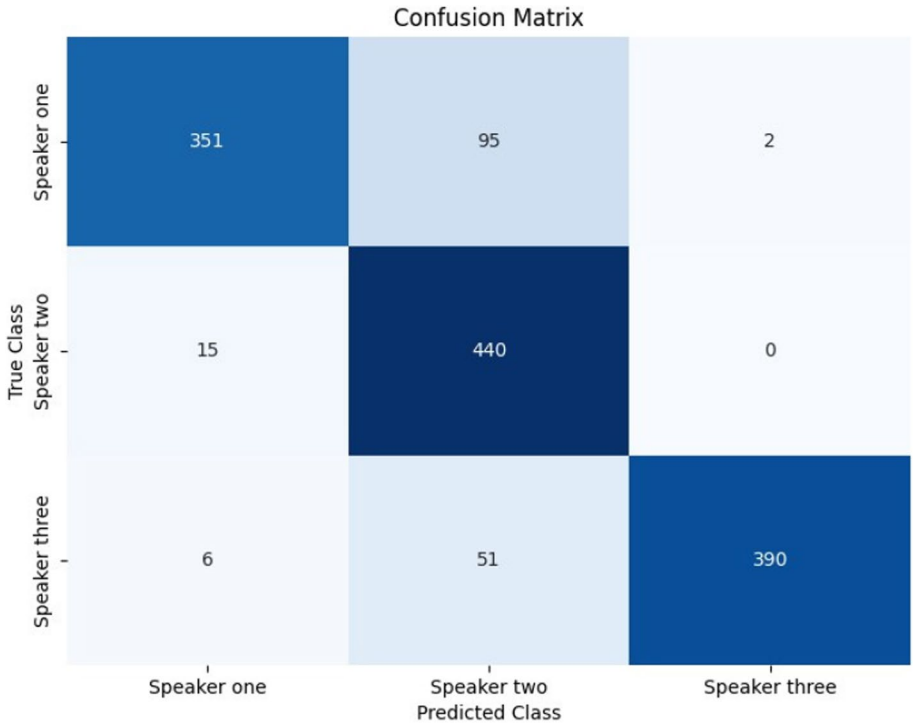

Fig. 11 Confusion matrix for case four

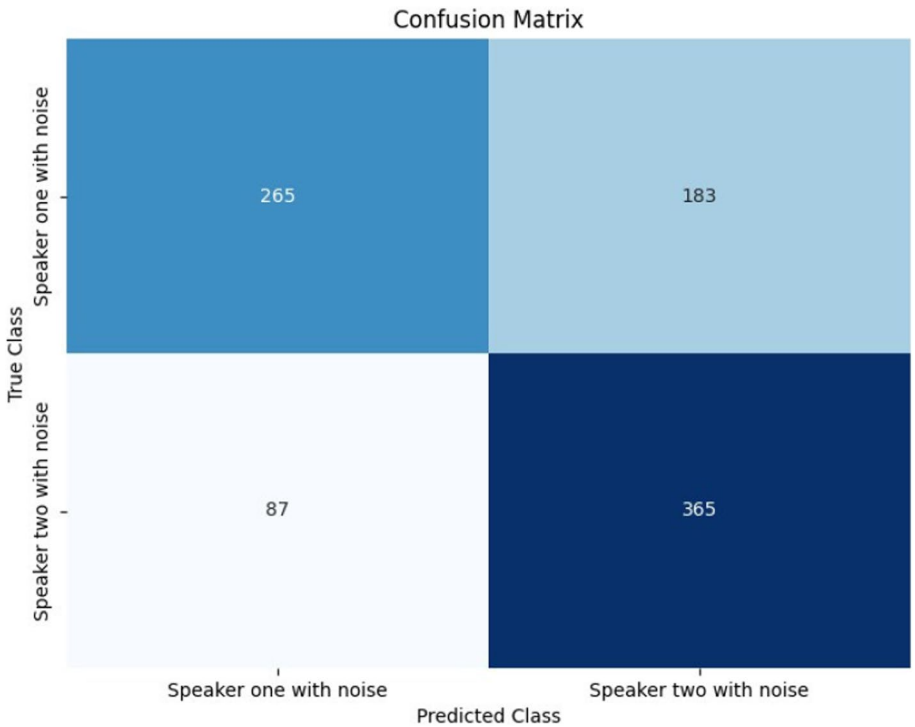

Fig. 12 Confusion matrix for case one with noise

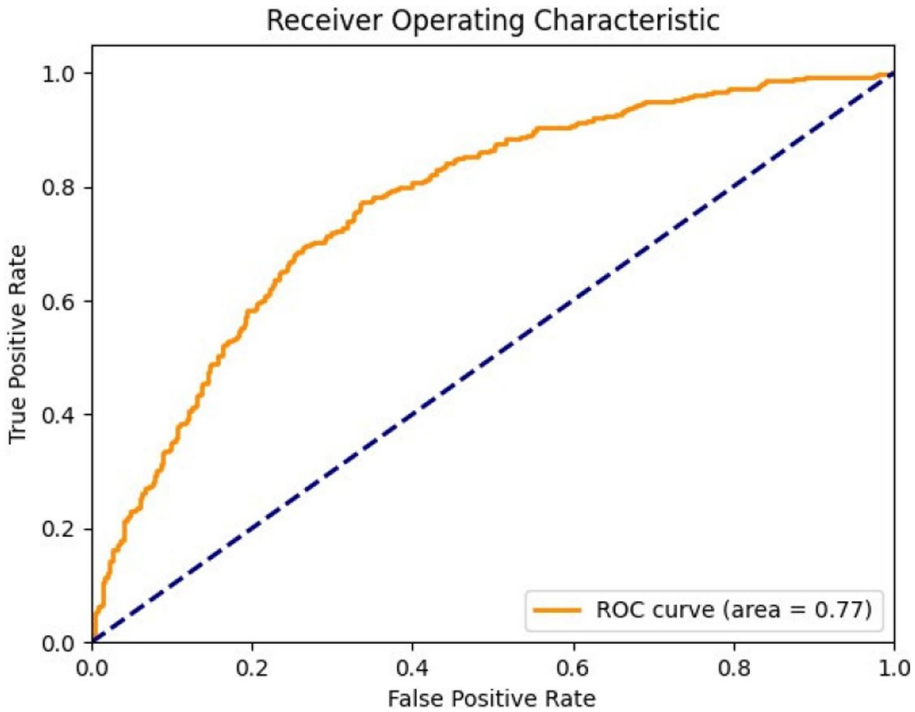

Fig. 13 Receiver operating characteristic for case one with noise

The categories 'Speaker one', 'Speaker two', and 'noise' were present in the second scenario. Of the 448 instances of this category in the Speaker 1 samples, 265 were correctly recognised, while 183 examples were wrongly classified. 365 out of 452 instances of this category in Speaker 2 samples were correctly recognised, while 87 examples were wrongly classified.

The confusion matrix demonstrates how the system's operation was hampered by the background noise. There was a decrease in accuracy and an increase in misclassifications as compared to the condition without noise. The number of misclassifications in each category shows how the system's ability to correctly identify the speakers is impacted by background noise.

The ROC area of 0.77 indicated that the system's discriminating abilities were still good, but somewhat lower than in the noise-free condition. It implies that background noise may cause the system to categorize things into different groups with slightly less precision.

The confusion matrix and ROC of identifying speaker sounds in the presence of background noise, 'Speaker one', 'Speaker three', and noise, are displayed in Figs. 14 and 15, respectively:

Three hundred thirty-six out of 444 samples in the second scenario, which included background noise and the category 'Speaker one', were correctly recognised, whereas 108 examples were wrongly classified. In the Speaker three category, 402 out of 260 instances had the right identification. 54 were misclassified, nevertheless. The ROC

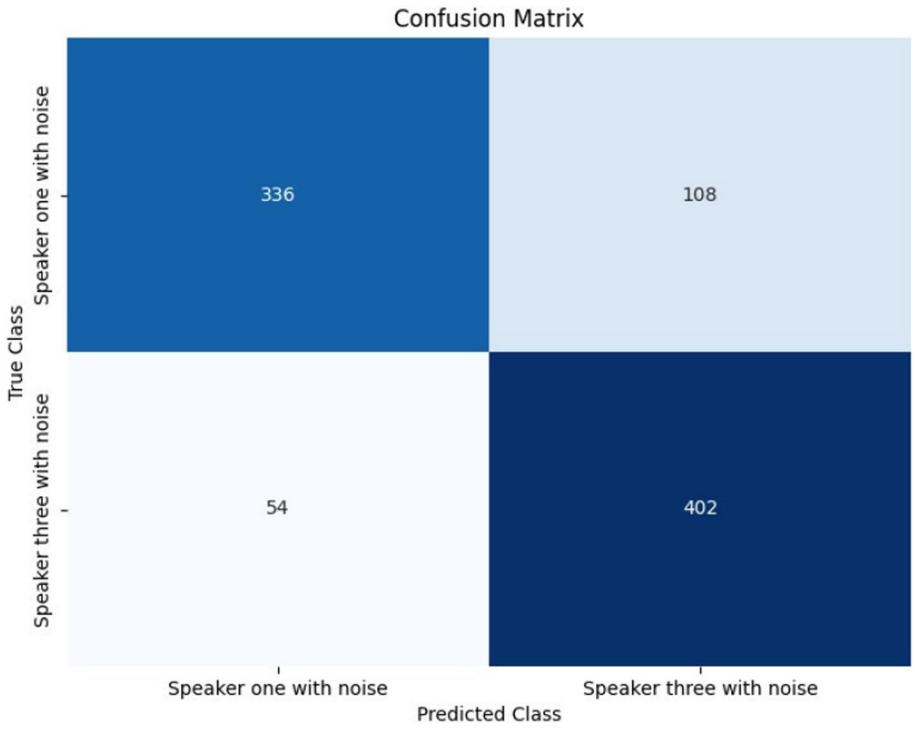

Fig. 14 Confusion matrix for case two with noise

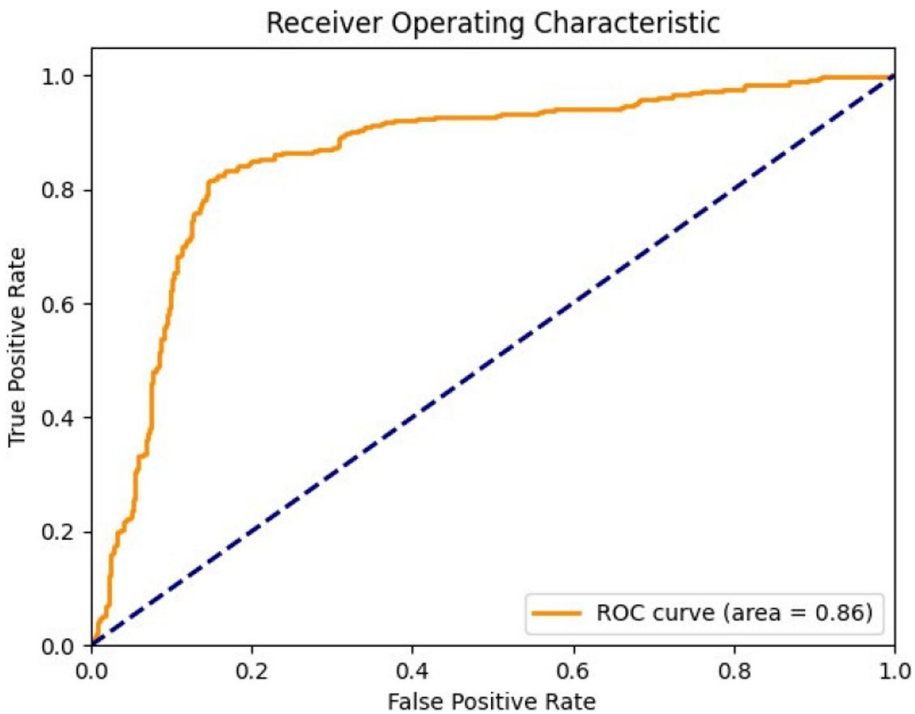

Fig. 15 Receiver operating characteristic for case two with noise

plot displayed an area of 0.86 despite background noise, indicating that the system had a high discriminating capacity to separate the two groups.

The confusion matrix demonstrates how the system's operation was hampered by the background noise. There was a decrease in accuracy and an increase in misclassifications as compared to the condition without noise.

The amount of false positives for every category shows how background noise degrades system performance. The ROC area of 0.86 indicates that the system had strong classification abilities, but much less than in the noise-free situation. It means that even in the presence of background noise, the system is still able to distinguish between the various categories.

The confusion matrix and receiver operating characteristic curve (ROC) for the speaker sounds ['Speaker three', 'Speaker two', and noise], respectively, are displayed in Figs. 16 and 17.

There are three categories in the third case study: 'Speaker three', 'Speaker two', and 'noise'. 387 of the 443 instances in the Speaker 2 category were correctly recognised, while 56 were incorrectly categorized.

In the Speaker three category, 328 instances out of 458 were accurately recognised. But 129 were incorrectly categorized. The ROC plot displayed an area of 0.83 despite background noise, indicating the system's strong ability to discriminate between the two groups.

The confusion matrix demonstrates how the system's operation was impeded by the background noise. There was a decrease in accuracy and an increase in misclassifications as compared to the condition without noise. The amount of false positives for every category shows how background noise degrades system performance.

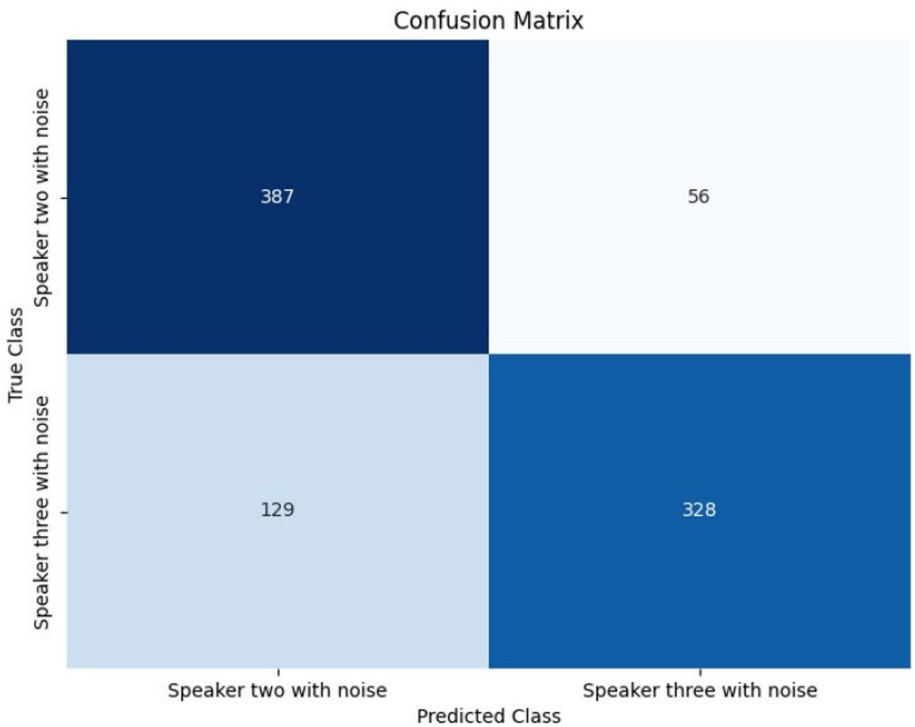

Fig. 16 Confusion matrix for case three with noise

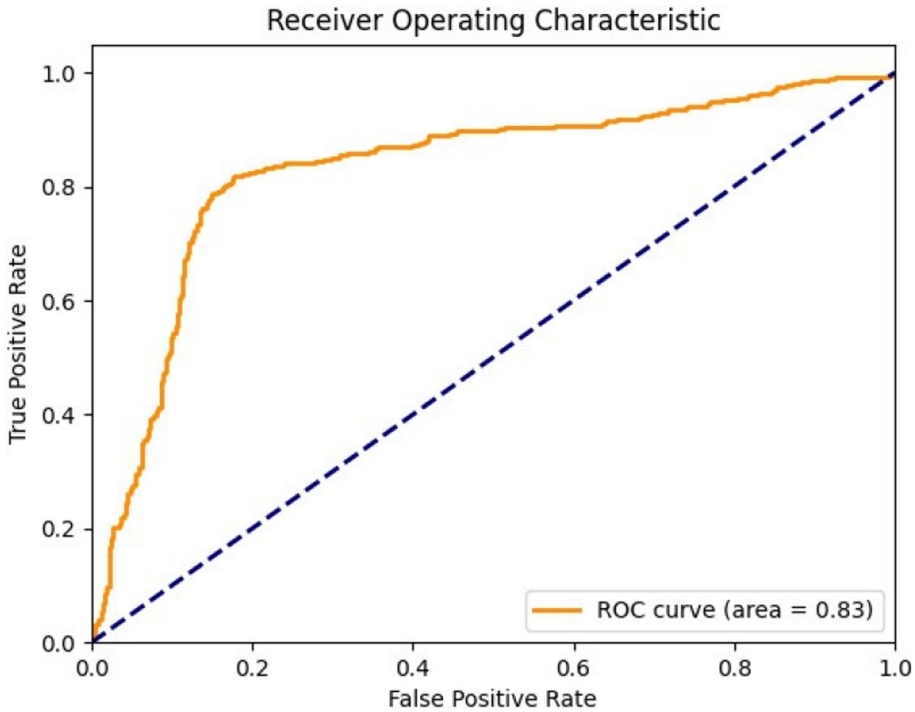

Fig. 17 Receiver operating characteristic for case three with noise

The program demonstrated high accuracy when classifying instances for the Speaker two and Speaker three categories, despite the presence of background noise.

The confusion matrix for categorising speaker sounds of 'Speaker three,' 'Speaker two', 'Speaker one', and noise is displayed in Fig. 18:

In the fourth case, which included background noise and the categories 'Speaker three,' 'Speaker two', and 'Speaker one', 321 of the 444 instances of the Speaker one category were properly identified, while 123 were incorrectly categorised.

Out of 456 instances in the Speaker two category, 246 were correctly identified, but 210 were misclassified. Of 450 instances in the Speaker three category, 318 were correctly identified. However, 132 were misclassified.

When there was no outside noise (Scenario One), the system did a remarkable job of differentiating between speakers. Case One's excellent accuracy and discriminating power were demonstrated by comparing 'Speaker one' with 'Speaker two', as evidenced by a ROC plot with an area under the curve (AUC) of 0.96. AUCs of 1 were seen in the ROC plots for Case Two ('Speaker one' vs. 'Speaker three') and Case Three ('Speaker three' vs. 'Speaker two'), which also demonstrated excellent classification accuracies. In contrast to the easier situations, Case Four, which had three speaker types, showed somewhat lower accuracy.

The system's performance was affected when Scenario two was switched to and background noise was introduced. In contrast to the noise-free environment, cases one, two, and three demonstrated a decline in accuracy and an increase in misclassifications. The system was nonetheless able to discriminate quite well in spite of these difficulties. AUC of 0.86 was shown in the ROC plot for case two, which comprised 'Speaker one', 'Speaker

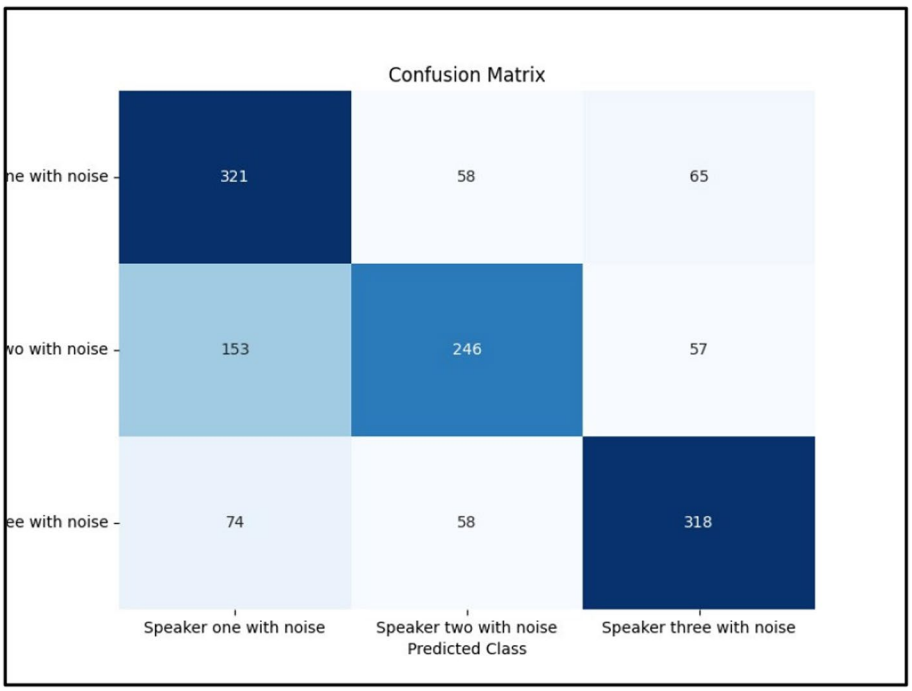

Fig. 18 Confusion matrix for case four with noise

three', and 'noise', demonstrating the system's capacity to distinguish between categories in the presence of noise.

With three speaker categories plus background noise in case four, the most complicated case, the system showed decreased accuracy and more misclassifications. It did, however, continue to show some capacity to distinguish between the different groups.

According to the overall summary, the developed system demonstrated great accuracy and discriminating power while operating extraordinarily well in perfect circumstances free from outside noise. Although the presence of background noise created difficulties and reduced accuracy, the system was still able to discern between various speaker sounds. This detailed assessment offers insightful information about how resilient the system is under different circumstances and presents a thorough picture of its practicality.

Conclusions

According to the summary, this research investigated for how recording technologies and environmental factors impact the precision and effectiveness of deep learning-based speech recognition systems. The initial stage in the paper's systematic approach to data collection was the collection of data from the speaker Recognition Dataset, which includes speakers from a variety of speakers. After the data's speaker audio was converted into spectrogram images, it was preprocessed and prepared for training a CNN model. The CNN model was trained using the spectrogram images, and its performance was evaluated using test samples.

An important step in evaluating the program's resilience and effectiveness is the careful division of the dataset into three similar voice groups according to voice tone

and gender. This division makes it possible to evaluate the software specifically in some areas, which helps to provide a more complex picture of its capabilities.

The study found that in situations without background noise, the CNN model was able to classify speakers with very high accuracy in several categories. But when background noise was present, the algorithm struggled to distinguish speakers accurately, especially for some categories. High AUC values in the ROC analysis and the results indicated that, even in the presence of background noise, the system continued to show excellent detection skills.

In accordance with the findings, it is advised to improve the speech recognition system's effectiveness in noisy environments even further. This may include looking at techniques like noise reduction algorithms or adding preprocessing steps to increase the system's ability to discriminate between background noise and voice input. Expanding the dataset to cover a wider range of speakers and different environmental conditions would offer more information on the system's reliability and performance in real-world scenarios.

Future research should focus on addressing the speaker kinds' misclassifications as well. Examining the unique characteristics of those speakers' voices and taking into account fresh components or models that can more accurately capture and distinguish them may be necessary to achieve this.

All things considered, this study provides valuable insights into how recording technologies and environmental factors impact the efficiency and reliability of deep learning-based speech identification systems. Through the resolution of identified problems and the expansion of the system's capabilities, these recommendations aim to improve the accuracy and reliability of speaker identification algorithms in practical applications.

Abbreviations

SI	Speaker identification
ASR	Automated speech recognition
DL	Deep learning
DSR	Deep speaker recognition
GMM	Gaussian mixture models
HMM	Hidden Markov models
UBM	Universal Background Models
CNNs	Convolutional neural networks
PCM	Pulse-code modulation
AUC	Area under the curve
TPR	True positive rate
FPR	False positive rate

Acknowledgements

We extend our sincere appreciation to all those who contributed to the successful completion of this research. In particular, we would like to acknowledge the invaluable teamwork and collaboration between the authors, Omar Ratib Khazaleh and Leen Ahmed Khrais.

The tests, data analysis, and methodology development described in this work were all made possible by their combined efforts, mutual support, and shared dedication. Our research's quality and depth were considerably improved by their spirit of cooperation and collective knowledge. We genuinely appreciate the teamwork and synergy that enabled this task to be completed.

Authors' contributions

Data preprocessing: Omar Ratib Khazaleh performed data preprocessing tasks, including data cleaning and formatting. Model training: Omar Ratib Khazaleh was involved in training the convolutional neural network (CNN) model. Data collection: Leen Ahmed Khrais was responsible for searching the audio data used in this study. Data analysis: Leen Ahmed Khrais conducted the initial analysis of the collected data, and contributed to the development of the research methodology. Performance evaluation: Both authors contributed to the evaluation of the model's performance.

Funding
This research received no specific funding.

Availability of data and materials
The 'Speaker Recognition Dataset', which is available at [24], is where the audio data for this investigation was obtained. These dataset's resources can be used by researchers and developers who are interested in speech analysis and identification.

Declarations

Competing interests
No conflicts of interest that might have affected the research or how the results were presented have been disclosed by the authors.

Received: 26 September 2023 Accepted: 20 December 2023
Published online: 06 January 2024

References

1. Jadhav S, Karpe S, Das S (2021) Sound classification using python. In: ITM Web of Conferences. EDP Sciences, vol. 40, p 03024
2. Mukhamadiyev A, Khujayarov I, Djuraev O, Cho J (2022) Automatic speech recognition method based on deep learning approaches for Uzbek language. Sensors 22(10):3683
3. Tuunanen T (2020) Real-time sound event detection with python. (Master's thesis)
4. Ohi A, Mridha MF, Hamid MA, Monowar MM (2021) Deep speaker recognition: process, progress, and challenges. IEEE Access 9:89619–89643
5. Le Q, Miralles-Pechuán L, Kulkarni S, Su J (2020) An overview of deep learning in industry. Data Anal AI 1:65–98
6. Sharma P, Abrol V, Sao AK (2017) Deep-sparse-representation-based features for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 25(11):2162–2175
7. Fadlullah Z, Tang F, Mao B, Kato N, Akashi O, Inoue T, Mizutani K (2017) State-of-the-art deep learning: evolving machine intelligence toward tomorrow's intelligent network traffic control systems. IEEE Commun Surv Tutor 19(4):2432–2455
8. Zhang J, Tao D (2020) Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. IEEE Internet Things J 8(10):7789–7817
9. Hansen J, Hasan T (2015) Speaker recognition by machines and humans: A tutorial review. IEEE Signal Process Mag 32(6):74–99
10. Gonzalez-Rodriguez J (2014) Evaluating automatic speaker recognition systems: An overview of the nist speaker recognition evaluations (1996–2014). Loquens 1(1):e007–e007
11. Hutiri W, Ding AY (2022) Bias in automated speaker recognition. In: 2022 ACM conference on fairness, accountability, and transparency, p 230–247
12. Jahangir R, Teh YW, Nweke HF, Mujtaba G, Al-Garadi M (2021) Speaker identification through artificial intelligence techniques: a comprehensive review and research challenges. Expert Syst Appl 171:114591
13. Abdullah H, Warren K, Bindschaedler V, Papernot N (2021) Sok: the faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems. In: 2021 IEEE symposium on security and privacy (SP), p 730–747
14. Hanifa R, Isa K, Mohamad S (2017) Malay speech recognition for different ethnic speakers: an exploratory study. In: 2017 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE). IEEE, Piscataway p 91–96
15. Mary L, Yegnanarayana B (2008) Extraction and representation of prosodic features for language and speaker recognition. Speech Commun 50(10):782–796
16. Hanifa R, Isa K (2021) A review on speaker recognition: technology and challenges. Comput Electr Eng 90:107005
17. Nolan F (1983) The phonetic bases of speaker recognition. Cambridge UP, Cambridge
18. Tirumala S, Shahamiri SR, Garhwal AS, Wang R (2017) Speaker identification features extraction methods: a systematic review. Expert Syst Appl 90:250–271
19. Tirumala S, Shahamiri SR (2016) A review on deep learning approaches in speaker identification. In: Proceedings of the 8th international conference on signal processing systems, p 142–147
20. Saquib Z, Salam N, Nair RP, Pandey N (2010) A survey on automatic speaker recognition systems. In: International conference on multimedia, computer graphics, and broadcasting, p 134–145
21. Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: from features to supervectors. Speech Commun 52(1):12–40
22. Bai Z, Zhang XL (2021) Speaker recognition based on deep learning: an overview. Neural Netw 140:65–99
23. Pawar R, Jalnekar RM, Chitode JS (2018) Review of various stages in speaker recognition system, performance measures and recognition toolkits. Analog Integr Circ Sig Process 94:247–257
24. Kaggle (2023) Speaker recognition dataset. Kaggle. Available: https://www.kaggle.com/datasets/kongaevans/speaker-recognition-dataset

Publisher's Note