Enhancing speaker verification accuracy with deep ensemble learning and inclusion of multifaceted demographic factors

Pranita Niraj Palsapure, Rajeswari, Sandeep Kumar Kempegowda
Department of Electronics and Communication Engineering, Acharya Institute of Technology, Bangalore, India

Article Info

Article history:

Received Feb 13, 2023
Revised May 12, 2023
Accepted Jun 4, 2023

Keywords:

Deep ensemble learning
Disparity mitigation
Speaker verification
Speech signal processing
VoxCeleb 2

ABSTRACT

Effective speaker identification is essential for achieving robust speaker recognition in real-world applications such as mobile devices, security, and entertainment while ensuring high accuracy. However, deep learning models trained on large datasets with diverse demographic and environmental factors may lead to increased misclassification and longer processing times. This study proposes incorporating ethnicity and gender information as critical parameters in a deep learning model to enhance accuracy. Two convolutional neural network (CNN) models classify gender and ethnicity, followed by a Siamese deep learning model trained with critical parameters and additional features for speaker verification. The proposed model was tested using the VoxCeleb 2 database, which includes over one million utterances from 6,112 celebrities. In an evaluation after 500 epochs, equal error rate (EER) and minimum decision cost function (minDCF) showed notable results, scoring 1.68 and 0.10, respectively. The proposed model outperforms existing deep learning models, demonstrating improved performance in terms of reduced misclassification errors and faster processing times.

Corresponding Author:

Pranita Niraj Palsapure
Department of Electronics and Communication Engineering, Acharya Institute of Technology
Bangalore, India
Email: pranitanirajpalsapure@gmail.com

1. INTRODUCTION

Using voice biometric technology based on automatic speaker recognition will revolutionize how we interact with technology in the near future [1]. It is widely used in various applications, including speaker recognition, voice-enabled smart devices, access control systems, and telephone banking systems. In these systems, a person's voice is combined into a single image used to authenticate or identify an individual [2]. One of the main advantages of voice biometrics is its non-intrusiveness. Unlike other biometric methods like fingerprint or facial recognition, voice biometrics can be performed remotely and does not require the user's presence. To create a unique voiceprint of a speaker, the technology analyzes their voice's spectral and prosodic characteristics, such as pitch, formants, and rhythm. Afterward, the speaker's voiceprint can be used to identify him or her in the future. Voice-based biometric systems include two main components: speaker identification and speaker verification [3]. In speaker identification systems, unknown speakers are identified by comparing their voices with a database of enrolled speakers. Speaker verification systems are used to confirm the identity of a speaker by comparing their voice to recorded voice samples in the database [4]. However, the existing systems are subject to inequality concerns, especially for those from different demographic groups. Apart from this, the existing automatic speaker recognition approaches are prone to disparity issues which arise when there are differences in system performance due to race, gender, accent, and age. Several factors contribute to these issues, such as biased datasets, voice variations, speaker characteristics sensitive to pronunciation, and algorithms that favor specific speakers [5]. Due to these disparities, erroneous results can be generated, or a high rate of false rejections occur, i.e., legitimate users are denied access or fraudulent approvals are granted, such as for certain groups of speakers, which can create a negative user experience and have security implications. To address the disparities in speaker recognition systems, researchers have proposed data augmentation and domain optimization with artificial intelligence (AI) [6], [7]. The researchers have also explored the ensemble mechanism, which combines different learning techniques to enhance classifier accuracy in various predictive tasks [8]. In the speaker recognition context, combining multiple biometric modalities, such as voice and facial features, has been suggested to enhance the authentication process [9], [10]. While adding extra features can improve the system's accuracy, it also brings with it an increased computational burden and higher resource requirements. This can result in slower recognition times and a less seamless user experience. Furthermore, these multi-modality systems can also be susceptible to background noise and external factors, affecting their performance.

This paper presents a highly integrated model that leverages the strengths of ensemble learning and considers ethnicity and gender features to produce a more robust and accurate prediction of speaker identity uniqueness. By addressing the challenges posed by complex features, speech, and variability, the proposed work reported in this paper has the potential to significantly improve the performance and user experience of speaker recognition applications. However, few research works have been done in a similar context considering text data and speech signals in the recent literature [11], [12]. Prachi *et al.* [13] emphasized designing a framework based on deep learning (DL) to identify a person considering biometric characteristics associated with the speaking voice samples. The study evaluated the model considering Texas Instruments Massachusetts Institute of Technology (TIMIT) and LibriSpeech datasets followed by closed-set and open-set methodology. The experimental result shows higher accuracy in the closed-set and relatively lower accuracy in the open-set method. Hajavi and Etemad [14] presented a customized learning model, UtterIdNet, to perform speaker verification using short-length audio samples from the VoxCeleb2 dataset. This model utilizes features obtained from individual hidden layers and fed to three fully connected networks. However, when tested in a noisy environment, the method may be sensitive to higher false positive results. Bunrit *et al.* [15] a multi-layer neural network is designed and trained from scratch for text-independent gender recognition, with each signal wave sample transformed into a spectrogram as input. The result shows an improvement of approximately 4% compared to the mel-frequency cepstral coefficient (MFCC) based method. India *et al.* [16] have presented a non-fixed length gender recognition system by integrating an attention model with a convolutional neural network (CNN). However, this work lacks effective benchmarking to justify the effectiveness of their presented system.

Alkhawaldeh [17] emphasizes the importance of evaluating several sets of audio features and various machine learning models to develop a better system to identify gender. The result shows that the recall rate obtained for CNN is 99.97% and for support vector machine (SVM) is 99.7%, and with an optimal feature selection, both models achieved a 100% recall rate. Greco *et al.* [18] focused on optimizing the tradeoff between the accuracy and speed of the CNN-based speaker verification system. Hamdi *et al.* [19] used the concept of an ensembled learning mechanism to automate feature extraction tasks for speaker recognition from Arabic speech. Alashban and Alotaibi [20] present a different work of gender classification from mono-language data. A bidirectional long short-term memory (Bi-LSTM) model is employed as a classifier, and the results show higher accuracy of 91.7% obtained for Arabic speakers and 86.53% for English speakers. Alaliyat *et al.* [21] tried to develop a robust speaker verification system using the joint approach of CNN and long short-term memory (LSTM) for the door access and control mechanism security application. Kanervisto *et al.* [22] focused on spoofing attacks on the automatic speaker verification system and presented an optimized security function based on the self-exploration capabilities of reinforcement learning. Peri *et al.* [23], the researchers aimed to handle gender biases and improve fairness through adversarial and multi-task learning techniques. Lin and Mak [24] have presented a deep-weight space ensemble technique for handling mixed text-dependent and text-independent speaker verification, cross-channel speaker verification, and bi-lingual speaker verification. Atenco *et al.* [25] suggested a bimodal multi-task model for an audiovisual system that combines facial traits and speech data for user identification. The method was evaluated using the BIOMEX-DB and VidTIMIT databases. Vanderreydt and Demuynck [26] have presented a technique for speech recognition by estimating channel characteristics through time-frequency speech patches. This method improves accuracy in noisy speech recognition and maintains performance in clean speech. But this technique may not perform well on other noise or channel conditions not tested in the study, especially if the speech data has different characteristics than the data used. The literature analysis suggests there is still room for improvement in developing a practical computational approach for a reliable speaker recognition and verification system.

The remaining part of this paper is described as follows: section 2 briefly discusses the research problem and its contribution, followed by the proposed solution and system design. Section 3 discusses the method and implementation procedures adopted in the proposed system for speaker gender classification, ethnicity classification, and speaker verification. Section 4 presents the outcome and performance evaluation to justify the scope of the proposed method. Finally, section 5 concludes the actual contribution of this paper.

2. THE PROBLEM DESCRIPTION

The analysis of existing literature shows that different researchers have proposed different solutions to improve speaker identification and verification. Despite progress in the field, substantial issues still need to be overcome. A significant challenge is aggregating features and capturing long-term dependencies from different biometric data. Conventional machine learning techniques have difficulty recognizing speakers in noisy backgrounds and may not be practical in environments with limited training data. In addition, these techniques encounter disparity issues when identifying speakers from diverse demographic backgrounds. These challenges have led researchers to adopt large and complex network configurations, increasing parameters and response times. As a result, most existing approaches are computationally inefficient, making them incompatible with real-time systems. The trend in the literature shows that research is still underway to develop robust and accurate methods that can be deployed in real-life scenarios. This factor motivates this research work to advance the automatic speaker verification system field by efficiently harnessing the power of deep-learning approaches using a shared representation of multi-architecture frameworks through an automated pipelines approach. All significant problems were considered, and a comprehensive speaker identification and verification system was built with a computationally efficient ensemble-based method. In particular, the proposed study addresses the following concerns: i) the conceptualization of speaker verification as an adjustment to language, tone, and dialect; ii) the capture of the representative features of all groups of speakers with different demographic backgrounds using web scraping technique; iii) reducing computational complexity by separating entire models into pipelines; and iv) synchronizing key parameters and responses in an end-to-end manner, making speaker verification models robust by fully utilizing embedding channel information, and time-frequency information.

2.1. The proposed solution

The research work reported in this paper is primarily focused on developing a practical approach to creating a sustainable speaker identification system and ensuring its usefulness in various real-world applications, including homeland security, forensics, security, surveillance, and access control. The proposed system has been designed to adapt to changes in a speaker's voice even if there is variability in speech or the person who identifies as female but has a male-sounding voice. In this situation, the system needs not go through the database of the men as it utilizes a graph database to control such errors. System design is lightweight because complex data processing operations and handcrafted feature engineering are unnecessary. Instead, it automates the feature extraction by incorporating web-scrapping, two independent 1-dimensional convolutional neural networks (1D-CNNs) [27], and a twin neural network (Siamese) to capture precise speaker attributes in the embedding layers [28]. The proposed speaker verification system is trained over large samples on multiple tasks, including gender classification and ethnicity identification. Their response vector is a shared representation for speaker verification, improving the system's overall performance and reducing disparities.

Figure 1 illustrates the block-based architectural design and workflow of the proposed deep learning ensemble-based speaker verification system, which comprises four core modules such as i) web scraping for data labeling, ii) gender classification, iii) ethnicity identification, and iv) speaker verification. Both ethnicity and gender classification models are implemented using 1D CNNs and trained independently. A Siamese network combines the responses of a trained gender and ethnicity classifier in a synchronized and end-to-end manner to detect and verify speakers with high objectivity.

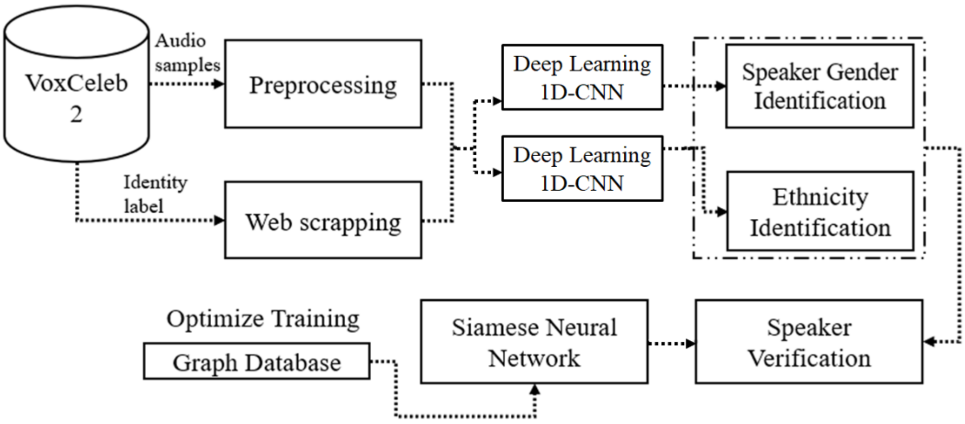

Figure 1. Proposed deep learning ensemble for speaker verification

3. METHOD

This section comprehensively discusses the implementation of the proposed system. First, the adopted dataset is briefly described, followed by extracting the data labels using a web scraping method. The second half of this section focuses on implementing the speaker classification module, which is used to identify the gender and race of the speaker. Further, an implementation strategy was adopted for verifying the speaker's identity discussed. Each module has a detailed explanation, including algorithmic steps and mathematical expressions used.

3.1. Dataset description

A VoxCeleb 2 dataset is used in the proposed work [29]. Researchers at the University of Oxford created VoxCeleb 2 to collect audio and video recordings of people speaking. It was designed for training and evaluation systems for speaker recognition, which identifies a person speaking in an audio or video recording. The dataset includes approximately 1.2 million utterances from 5,994 different speakers and a wide variety of accents, languages, and recording conditions. It is one of the largest and most diverse datasets and is widely used in research on speaker recognition and related tasks. The statistics of the VoxCeleb 2 dataset are tabulated in Table 1.

Table 1. An overview of VoxCeleb 2 dataset attributes and statistics

Data files	dev	Test
# of speakers	5,994	118
# of videos	145,569	4,911
# of utterances	1,092,009	36,237

3.2. Web scraping

The dataset VoxCeleb2 used in this study contains a wide variety of information about speakers, including audio samples, accent information, and gender information. The proposed work assumes that encoding speaker attributes can result in more descriptive features in the embedding layers of the learning model. This set of features is significant for the authentication or verifying the speaker's ethnicity following the gender class (male or female). The dataset does provide the gender along with the name of the file. Therefore, the study uses the web scraping mechanism to collect the identity in the form of structured labels as it provides the informative characterization of the speaker and allows the learning model to capture better the gender variation in its learning stage and deployment phase. Since the most repeated noun in the description data in VoxCeleb2 is the celebrity's name. Web scraping is done to get the person's or celebrity's name and racial identity. For this, each identity in the VoxCeleb2 is mapped to a video over YouTube, and each of the pages is grabbed, and the description is read. A computing step executing web scrapping for fetching speaker names from the video is illustrated in algorithm 1.

Algorithm 1. Web scrapping to get labels name of speaker
```
for each dataset → Tr (Training set), Ts (Testing set)
   get a list of folders within the specified VoxCeleb directory
   get a list of videos within the folder
   Init titles → [ ]
   for each video
      get html of video's YouTube page
       title=parse HTML using BeautifulSoup (Python library)
         extract the video title and convert it to lowercase
         append title → to titles list
         name=[ ]  // empty list to store the name of the person in the videos
         first=0
         second=0
       while the length of the name is not equal to 2:
        name=get common words in titles[first] and titles[second] using common() function
       increment first by 1
       If the first is equal to 4 and the second is equal to 4:
    exit loop
    If the first is equal to 4:
        first=0 // reset
        increment second by 1
      print joined version of the name
end
```

The computing steps in Algorithm 1 describe a web scraping process for a VoxCeleb dataset to get a label as a speaker name. The process starts by getting a list of folders within the specified VoxCeleb directory and a list of videos within each folder. Then, an empty list of “titles” is initialized to store the titles of the videos. For each video, the hypertext markup language (HTML) of the video's YouTube page is obtained, and the title is extracted using the BeautifulSoup library in Python. The extracted title is converted to lowercase and added to the “titles” list. Next, an empty list “name” is created to store the person's name in the videos. The first and second variables are initialized to 0. A while loop is then executed to get the common words in the titles of two videos using the “common()” function. The “first” variable is incremented by 1; if it reaches four and the “second” also reaches 4, the loop is exited. If “first” reaches 4, it is reset to 0, and “second” is incremented by 1. Finally, the joined version of the “name” list is printed. The pseudocode provides a general outline of the process for web scraping the VoxCeleb dataset and extracting the names of the people in the videos. Similar computing steps are adopted for fetching another label in the form of ethnicity identity. The following subsection elaborates on implementing the speaker gender classification model trained using labels obtained from web scraping.

3.3. Speaker gender classification

This phase of the proposed system executes the speaker identification process based on its gender using characteristics of their voice such as pitch and formants. This phase's response serves as an additional factor to improve the performance of the speaker verification task. For example, a system may require a match not only in the speaker's voice but also in their gender. This way, the system can be more robust to voice variations due to different accents, dialects, and speaking styles. Identification of the speaker's gender usually occurs through comparing particular utterances with templates. The study has implemented multi-layer neural network architectures with a 1D CNN unit to deal with this situation. Before feeding the whole feature vector into the network, the input speech data is preprocessed. The system first loads the speech signal Sig and converts into an array, which is then normalized to ensure that the values are within a specific range, as given in (1).

$$Sig_N = \frac{(Sig - min\ (Sig))}{max(Sig) - min\ (Sig)} \tag{1}$$

The process of signal normalization, as shown in (1), will help improve the model's training stability. The next step of the preprocessing executes the windowing Win operation over the normalized signal Sig_N to split into a temporal segment, numerically given in (2),

$$w(n) = \alpha - (1 - \alpha) \times cos\left(\frac{2\pi n}{T-1}\right) \tag{2}$$

where, *w(n)* is a mathematical function of Win, α refers to the coefficient of Win, n denotes frame size ranging from 0 to $N - 1$ and T is the total length of Sig. The window function is applied to each frame to reduce the effects of spectral leakage, which can occur when a signal is analyzed using a finite-length window, numerically expressed in (3),

$$Sig\_length = (win - 1)x(win\_step) + (win\_length) \tag{3}$$

where win_step is 500 ms, and win_length equals 1000 ms. This process reduces the effects of spectral leakage by analyzing short portions of a long signal and its frequency content, thereby providing a fast convergence rate to the learning model.

In the next operation, the feature vector attained after normalization and windowing operation merges with the binary label obtained through the web scrapping process. The obtained class label is binary because the objective is to categorize the audio stream into male or female. It is to be noted that no additional features are extracted before training the model. The study uses 1D-CNN, similar to a 2-D CNN, but instead of working with 2-D arrays, a 1-D CNN works with 1-D arrays of data. Therefore, it can be trained to learn the speaker attributes directly from the preprocessed speech signals. However, in some cases, features such as MFCCs are extracted to reduce the data dimensionality and make the training process faster. But it also depends on the nature of the problem and since we are dealing with a gender classification model for a real-time system that receives complex and variable voice signals. The more features the model is exposed to, the better it learns. The CNN model implemented in the study consists of an input layer, 4 number of 1D convolutional layers along, with a max pooling layer and a fully connected output layer. The convolutional layers use filters to extract features such as pitch and frequency information from the input signal Sig. The convolution operation is applied with a sliding window, where the filter moves across the input data, element-wise multiplying the filter values with the corresponding input data values and summing them up to obtain the output feature map, as given in (4).

$$y(i) = \sum w(k) \times Sig(i + k) + b \quad (4)$$

where, $y(i)$ denotes output feature map of the i^{th} element of input Sig, $w(k)$ refers to the kernel or convolutional filter, b is the bias parameter, and $\sum$ denotes the summation operation. The following component of the convolution operation is the max pooling which reduces the dimensionality of the feature maps produced by the convolutional layers while retaining important information. This operation helps reduce overfitting and improve computational efficiency, numerically given in (5).

$$y(i) = \max\left(Sig(i{:}\, i + k)\right) \quad (5)$$

Finally, the output layer with a single neuron and activation function sigmoid predicts the gender of the speaker from the input voice signal. This function is commonly used for binary classification problems as it provides a probabilistic interpretation of the output. In this case, the sigmoid function is used to obtain a value between 0 and 1, representing the probability of the speaker being male or female.

3.4. Ethnicity classification

Similar to speaker gender identification or classification, a speaker's race can be used to verify a speaker's identity. In previous works, this process is also called speaker binarization. In this process, various acoustic features of the speech, such as accent and articulation, are used to train the model and identify the speaker's ethnicity. Unlike existing studies, the proposed study does not use the input as a speech signal. Instead, it uses ethnic labels corresponding to speaker identity, gender, and person's name, obtained through a web scraping method. The uniqueness of this method is that it does not involve any complicated mechanism since it does not need to process the speech signal and extract complex features. Instead, it focuses on predicting a speaker's ethnicity and mapping it to a corresponding identity, including a person's name and gender, for the speaker verification process discussed in the next section. In this way, the proposed study ensures its validity and should be accurate enough to identify speakers in the presence of subgroups of people with a common cultural background. Ethnicity recognition is done with 1D-CNN, following similar network configurations used for gender classification.

Figure 2 shows that the dataset created is associated with a class imbalance factor, which may make the learning model biased towards the majority class. Therefore, the study uses a Regularizer mechanism in order to calibrate 1D-CNN and control the biases in the learning process by guiding what to learn and what not to learn. This Regularizer mechanism is developed using a sequential model with a linear stack of neural layers. The first layer has 1,000 neuron units, the second has 100 neurons, the third has ten neurons, and the fourth has one neuron. This Regularizer network is deployed at the hidden layer of the 1D-CNN; it takes the output from the hidden layer, learns the possible bias value, and gives its response value to the last layer of 1D-CNN to avoid biases in the feature generalization for the majority classes. The next module of the proposed system is to carry out the verification of gender, which is crucial in the application of gender recognition. In this phase, a Siamese neural network is implemented, and a graph database (DB) mechanism is used to avoid misleading factors in the prediction of gender by the learning model.

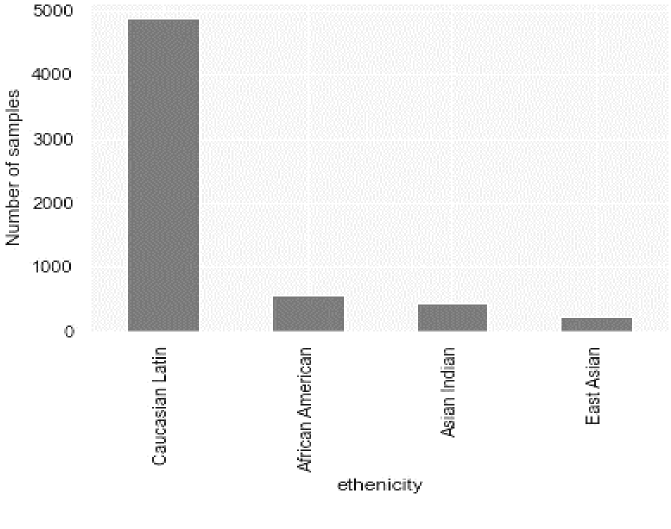

Figure 2. Ethnicity data distribution

3.5. Speaker verification

This implementation phase is carried out to determine whether a speaker's voice is authentic and belongs to a specific person. The output obtained from the previous models is pipelined to improve the accuracy with additional information about the speaker. By using gender and race classification, the system can be made more robust to variations in voices due to different accents, dialects, and speaking styles. The study uses the concept of a Siamese twin neural network, as shown in Figure 3. The learning model consists of two identical sub-networks, or “twins,” which share the same architecture and weights. The two sub-networks are trained together by providing them with pairs of inputs, where one input is a reference, and the other is a query. The network aims to learn a similarity function that can compare the two inputs and determine whether they are the same or different.

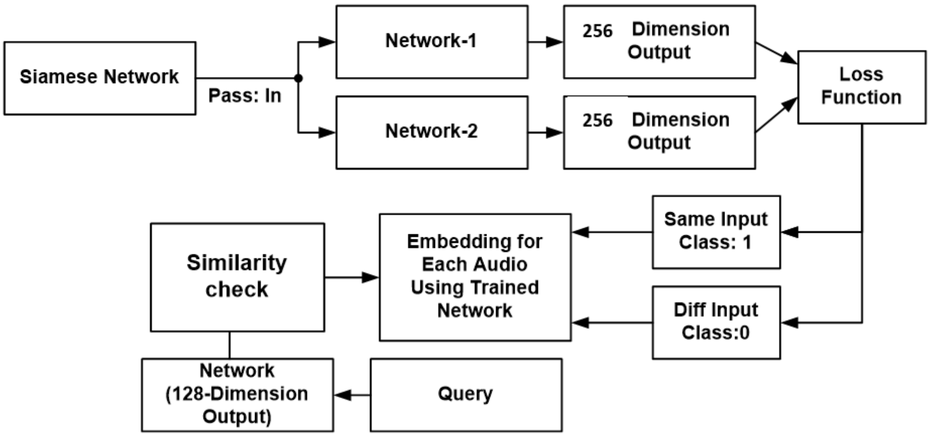

Figure 3. Speaker verification model using twin neural network

As shown in Figure 3, the speaker verification using a Siamese network is achieved by combining the outputs of a gender classifier and ethnicity classifier network, i.e., by leveraging the knowledge learned by the pre-trained CNNs for gender and ethnicity classification. Initially, the features are extracted by applying basic preprocessing operations over input voice samples. The extracted features are passed through the gender and ethnicity classifier networks to get the corresponding class probabilities which are then concatenated and used as input to a Siamese network. The Siamese network is trained to compare the concatenated class probabilities of a pair of speech audio to determine whether they are from the same speaker. The final output of the Siamese network is a similarity score between 0 and 1, where a score closes to 1 indicates that the speech audio is from the same speaker, and a score close to 0 indicates that they are not. In this way, by taking the output of the gender and ethnicity classifier networks as inputs to the Siamese network, the speaker verification system considers both the speaker's gender and ethnicity in the verification process, which can help improve the accuracy of the system.

The mathematical expression that summarizes the speaker verification procedure using a Siamese network: Let $X1$ and $X2$ be the speech audio input for speaker verification. Let $G1$ (male) and $G2$ (female) be the corresponding gender class probabilities obtained by passing $X1$ and $X2$ through the gender classifier network. Similarly, let us consider $E1$ and $E2$ to be the corresponding ethnicity class probabilities obtained by passing $X1$ and $X2$ through the ethnicity classifier network. Then the concatenated class probabilities in (6).

$$F1 = [G1, E1] \text{ and } F2 = [G2, E2] \tag{6}$$

The Siamese network has two branches; each takes one of the concatenated class probabilities as input and computes a high-dimensional embedding representation as numerically expressed in (7).

$$h1 = f(F1) \;\&\; h2 = f(F2) \tag{7}$$

where, $f(\cdot)$ is the embedding function. The similarity score between the two speech audio inputs is computed as the cosine similarity between the two embedding representations, mathematically expressed in (8).

$$S = f_1(h1, h2) = (h1 \cdot h2)/(||h1|| \cdot ||h2||) \tag{8}$$

where $f_1(\cdot)$ denotes a cosine similarity function, "." represents the dot product, and "‖.‖" represents the Euclidean norm.

The final output of the Siamese network is the similarity score S, which ranges between 0 and 1. A score of 1 indicates that the speech audio is from the same speaker, and a score of 0 indicates that they are not. The twin networks in the Siamese network are trained together by minimizing a triplet loss function, which measures the difference between the similarity score and the true label, such as 1 for similar inputs and 0 for different inputs. The goal of the loss function is to enforce the constraint that the distance between features of the same speaker should be smaller than the distance between features of different speakers. However, the learning of the model can be prone to misleading in the generalization of gender due to similar sound pitch or tonal distribution. For example, if a person's voice is soft, he might get classified as female even if he is male, or the reverse is possible as well as female with a complex range of frequency, then she may be predicted as male. Therefore, the proposed study uses a graph DB to handle the error by the artificial neural network (ANN), which may predict gender that may not be accurate. A graph database stores data in the form of nodes (neurons) and edges (artificial synapses or learnable parameters) that connect them for handling errors in a neural network. Storing the network's architecture in a graph database makes it easier to visualize and understand the network's structure, which can be helpful when identifying and troubleshooting errors. Figure 4 illustrates a sample visualization of graph DB.

As shown in Figure 4, the proposed mechanism of graph DB stores the network structure of the Siamese network, including the architecture of the two branches for gender and ethnicity classification. This process allows for easier visualization and understanding of the network's behavior, which can help identify and troubleshoot errors. On the other hand, the graph DB stores all the confidence values of the recognition for both gender and ethnicity. If the confidence value is low, the graph DB compares both male and female voices. This information is used to identify cases where the network makes predictions with low confidence, which may indicate an error in the model or the data. The advantage of this mechanism is that it offers an effective way to visualize and understand the network's structure, store confidence values, compare predictions, and track progress. On the other hand, Graph databases can help improve these models' accuracy and make them more reliable.

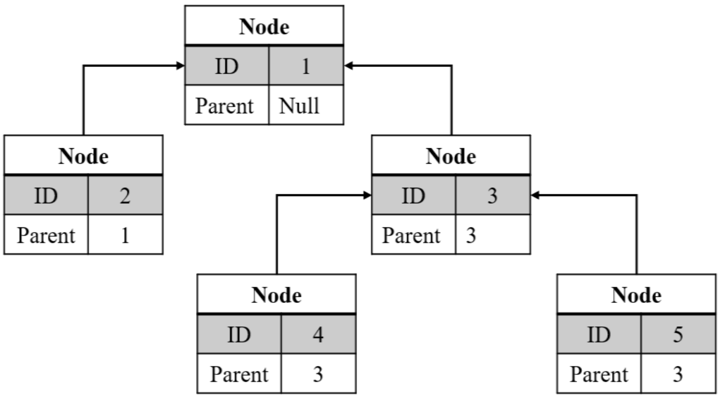

Figure 4. Illustration of graph DB for handling errors in network

4. RESULT AND DISCUSSION

The proposed system's performance is evaluated and analyzed for each module implemented in the Anaconda distribution using Python programming language. The performance analysis results are presented in this section, providing a clear insight into how well the system is working and where there is room for improvement. This involved a detailed examination of the output produced by the system and how well it performed in terms of accuracy, speed, and other relevant metrics. The performance analysis presented in this section provides valuable information about the system's strengths and limitations, helping to guide future development and improvements.

4.1. Performance analysis of gender classification model

Figure 5 shows the result of the proposed system for gender classification concerning the number of correct and incorrect predictions made by the classifier. The interpretation of the confusion plot is shown in Table 2. The confusion scores in Table 2 refer to the results of a gender classification model trained on a dataset with 3,700 samples of male speech and 2,500 samples of female speech. The scores indicate the number of true positive (TP) and false negative (FN) predictions made by the model for each gender. The model made

3,600 true positive predictions for male speech, meaning that it correctly classified 3,600 samples of male speech as male. However, the model also made 100 false negative predictions. On the other hand, the model did not make any errors in classifying female speech as male. The overall result indicates a high number of true-positive predictions showing no bias in the model towards both male and female speech. In this case, the model could be retrained with a more balanced dataset, or the network architecture could be modified to capture male speech characteristics better.

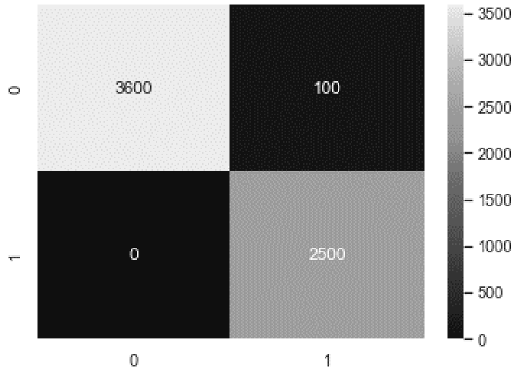

Figure 5. Confusion plot for gender classification

Table 2. Summary of confusion plot with classification statistics for gender classification

Total samples	Label	True positive	False negative
$Male = 3{,}700$	0	3,600	100
$Female\ = 2{,}500$	1	2,500	0

4.2. Performance analysis for ethnicity classification model

Figure 6 shows a multiclass confusion matrix for the 1D-CNN model, being evaluated on how well it can classify different ethnicities concerning the true positives and false negatives results. The ethnicity labels are encoded in the form of 0, 1, 2, and 3. For each ethnicity label, the TP count is shown for each ethnicity that the model classified correctly. The FN count is shown for each race that the model typed incorrectly.

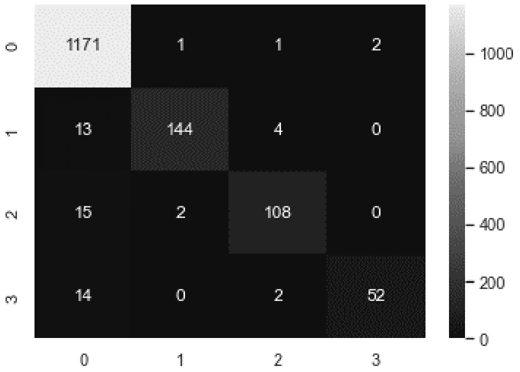

Figure 6. Confusion plot for ethnicity classification

The interpretation of this confusion plot is mentioned in Table 3. The scores show the performance of a classifier in correctly identifying different ethnicities. The classifier has four possible outputs corresponding to the four ethnicity labels: 0 (Caucasian Latin), 1 (African American), 2 (Asian Indian), and 3 (East Asian).

The TP scores indicate the number of instances correctly classified as belonging to each ethnicity. For example, it can be seen that the classifier correctly classified 1,171 instances as belonging to the Caucasian Latin ethnicity. The false negative FN scores indicate the number of instances misclassified as belonging to a different race. For example, there was one instance that was African American but was classified as Caucasian Latin. The results show that the classifier performs best for the Caucasian Latin ethnicity with a high true positive score and low false negative score. In contrast, the performance for the other ethnicities is not as bad, as evidenced by the false negative scores for these labels.

Table 3. Summary of confusion plot with classification statistics for ethnicity classification

Ethnicity labels	True positive	False negative		
Label=0	Caucasian Latin	African American	Asian Indian	East Asian
Caucasian Latin	1171	1	1	2
Label=1	African American	Caucasian Latin	Asian Indian	East Asian
African American	144	13	4	0
Label=2	Asian Indian	Caucasian Latin	African American	East Asian
Asian Indian	108	15	2	0
Label=3	East Asian	Caucasian Latin	African American	Asian Indian
East Asian	52	14	0	2

4.3. Performance analysis for speaker verification model

The performance analysis of the proposed model for speaker verification is carried out in terms of equal error rate (EER) and minimum decision cost function (minDCF) [30]. EER is a performance metric representing the point where the false acceptance rate (FAR) and false rejection rate (FRR) are equal. The EER is a critical metric because it balances the tradeoff between false rejections and false acceptances and provides a single number that summarizes the system's performance. The minDCF represents the minimum average cost per decision made by the system and provides a more comprehensive evaluation of the system's performance. The minDCF is calculated by weighting the FAR and false rejection rate FRR according to each error type's cost. The performance of the proposed speaker verification system is compared with the existing work proposed by Yao *et al.* [30] in terms of both EER and minDCF. Yao *et al.* [30] presented a speaker verification system based on the temporal embeddings obtained from multiple streams. The CNN classification model is implemented and validated on the same dataset adopted in our work. Table 4 illustrates the numerical value and comparative analysis for multistream CNN [30] and the proposed speaker verification method.

The numerical outcomes in terms of minDCF and EER measure over different epochs are shown in Table 4. The quantitative analysis reveals that the proposed speaker verification method remarkably improves overall training epochs, while the existing multistream CNN exhibits a less significant improvement. The comparison can be visualized in Figures 7 and 8. The graph trend demonstrates that the proposed system outperforms the existing model in terms of both EER with a 5.9% reduction and minDCF with a 6.1% reduction. The overall performance analysis suggests that the proposed system provides more accurate results in speaker verification than the existing system.

Table 4. Numerical outcome of a speaker verification system

Epochs	Multistream CNN [30]		Proposed deep learning ensemble	
	EER	minDCF	EER	minDCF
100	2.303	0.1609	2.41234	0.1742
200	1.919	0.1333	1.71532	0.1586
300	1.866	0.1381	1.70212	0.1286
400	1.834	0.1336	1.69124	0.1134
500	1.781	0.1310	1.68123	0.1032

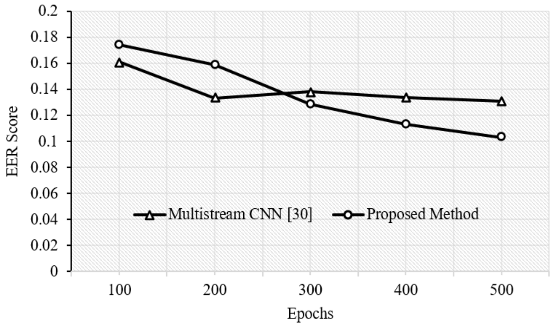

Figure 7. Proposed deep learning ensemble Vs. Multistream CNN [30] in terms of EER

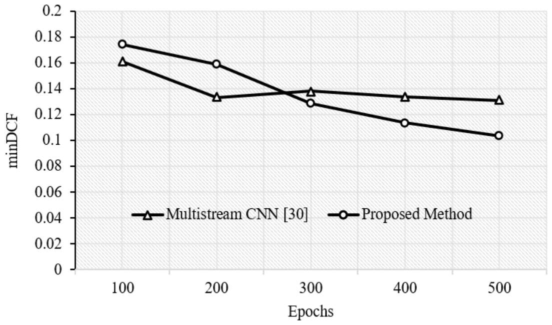

Figure 8. Proposed deep learning ensemble Vs. Multistream CNN [30] in terms of minDCF

5. CONCLUSION

This paper has introduced automated speaker verification system based on a deep ensemble learning mechanism addressing misclassification problem in diverse speaker demographic environments, and improving runtime efficiency when dealing with large datasets. The design of the formulated system involves web scrapping to extract precise labels from identity-tagged audio files stored in a database. Further, 1D CNN models are implemented to classify gender and ethnicity, and the outputs are utilized in a twin neural network model for speaker verification. In addition, a graph database is formulated to reduce misclassification errors. Finally, the twin neural network model with demographic parameters is trained for the speaker verification task. The proposed model shows significant improvement in terms of EER and minDCF compared to the existing models, while both were tested on the same VoxCeleb 2 database. Based on the simulation outcome, the proposed system's practicality is justified and can apply to real-world applications such as voice-enabled smart door system and access control systems. In future work, the proposed system will be extended to spoofing detection system by incorporating advanced learning and feature engineering operations.

REFERENCES

[1] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities," *IEEE Access*, vol. 9, pp. 79236–79263, 2021, doi: 10.1109/ACCESS.2021.3084299.

[2] L. Zheng, J. Li, M. Sun, X. Zhang, and T. F. Zheng, "When automatic voice disguise meets automatic speaker verification," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 824–837, 2021, doi: 10.1109/TIFS.2020.3023818.

[3] F. Abakarim and A. Abenaou, "Comparative study to realize an automatic speaker recognition system," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 1, pp. 376–382, Feb. 2022, doi: 10.11591/ijece.v12i1.pp376-382.

[4] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and antispoofing in the i-vector space," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 821–832, Apr. 2015, doi: 10.1109/TIFS.2015.2407362.

[5] A. Koenecke *et al.*, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, Apr. 2020, doi: 10.1073/pnas.1915768117.

[6] S. Singh, "Bayesian distance metric learning and its application in automatic speaker recognition systems," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 4, pp. 2960–2967, Aug. 2019, doi: 10.11591/ijece.v9i4.pp2960-2967.

[7] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications*, vol. 171, Jun. 2021, doi: 10.1016/j.eswa.2021.114591.

[8] S. K. Kempegowda, R. Rajeswari, L. Satyanarayana, and S. M. Basavarajaiah, "Hybrid features and ensembles of convolution neural networks for weed detection," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 6, pp. 6756–6767, Dec. 2022, doi: 10.11591/ijece.v12i6.pp6756-6767.

[9] D. Cai, W. Wang, and M. Li, "Incorporating visual information in audio based self-supervised speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1422–1435, 2022, doi: 10.1109/TASLP.2022.3162078.

[10] D. Watt *et al.*, "Assessing the effects of accent-mismatched reference population databases on the performance of an automatic speaker recognition system," *International Journal of Speech Language and the Law*, vol. 27, no. 1, pp. 1–34, Aug. 2020, doi: 10.1558/ijsll.41466.

[11] R. Mohd Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," *Computers and Electrical Engineering*, vol. 90, Mar. 2021, doi: 10.1016/j.compeleceng.2021.107005.

[12] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, May 2014, doi: 10.1016/j.specom.2014.03.001.

[13] N. N. Prachi, F. M. Nahiyan, M. Habibullah, and R. Khan, "Deep learning based speaker recognition system with CNN and LSTM techniques," in *2022 Interdisciplinary Research in Technology and Management (IRTM)*, Feb. 2022, pp. 1–6, doi: 10.1109/IRTM54583.2022.9791766.

[14] A. Hajavi and A. Etemad, "A deep neural network for short-segment speaker recognition," *Prepr. arXiv.1907.10420*, Jul. 2019.

[15] S. Bunrit, T. Inkian, N. Kerdprasop, and K. Kerdprasop, "Text-independent speaker identification using deep learning model of convolution neural network," *International Journal of Machine Learning and Computing*, vol. 9, no. 2, pp. 143–148, Apr. 2019, doi: 10.18178/ijmlc.2019.9.2.778.

[16] M. India, P. Safari, and J. Hernando, "Self multi-head attention for speaker recognition," *Prepr. arXiv.1906.09890*, Jun. 2019.

[17] R. S. Alkhawaldeh, "DGR: Gender recognition of human speech using one-dimensional conventional neural network," *Scientific Programming*, vol. 2019, pp. 1–12, Sep. 2019, doi: 10.1155/2019/7213717.

[18] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff," *IEEE Access*, vol. 8, pp. 130771–130781, 2020, doi: 10.1109/ACCESS.2020.3008793.

[19] S. Hamdi, A. Moussaoui, M. Oussalah, and M. Saidi, "Gender identification from Arabic speech using machine learning," in *MISC 2020: Modelling and Implementation of Complex Systems*, 2021, pp. 149–162, doi: 10.1007/978-3-030-58861-8_11.

[20] A. A. Alashban and Y. A. Alotaibi, "Speaker gender classification in mono-language and cross-language using BLSTM network," in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, Jul. 2021, pp. 66–71, doi: 10.1109/TSP52935.2021.9522623.

[21] S. Alaliyat, F. F. Waaler, K. Dyvik, R. Oucheikh, and I. Hameed, "Speaker verification using machine learning for door access control systems," in *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2021)*, 2021, pp. 689–700, doi: 10.1007/978-3-030-76346-6_61.

[22] A. Kanervisto, V. Hautamaki, T. Kinnunen, and J. Yamagishi, "Optimizing tandem speaker verification and anti-spoofing systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 477–488, 2022, doi: 10.1109/TASLP.2021.3138681.

[23] R. Peri, K. Somandepalli, and S. Narayanan, "A study of bias mitigation strategies for speaker recognition," *Computer Speech and Language*, vol. 79, Apr. 2023, doi: 10.1016/j.csl.2022.101481.

[24] W. Lin and M.-W. Mak, “Robust speaker verification using deep weight space ensemble,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 802–812, 2023, doi: 10.1109/TASLP.2022.3233231.

[25] J. C. Atenco, J. C. Moreno, and J. M. Ramirez, “Audiovisual biometric network with deep feature fusion for identification and text prompted verification,” *Algorithms*, vol. 16, no. 2, Jan. 2023, doi: 10.3390/a16020066.

[26] G. Vanderreydt and K. Demuynck, “A novel channel estimate for noise robust speech recognition,” *SSRN*, pp. 1–22, 2023.

[27] W. Tang, G. Long, L. Liu, T. Zhou, J. Jiang, and M. Blumenstein, “Rethinking 1D-CNN for time series classification: A stronger baseline,” *Prepr. arXiv.2002.10061*, Feb. 2020.

[28] S. Jadon and A. A. Srinivasan, “Improving siamese networks for one-shot learning using kernel-based activation functions,” in *Data Management, Analytics and Innovation*, 2021, pp. 353–367, doi: 10.1007/978-981-15-5619-7_25.

[29] A. Nagrani *et al.*, “VoxSRC 2020: The second VoxCeleb speaker recognition challenge,” *Prepr. arXiv.2012.06867*, Dec. 2020.

[30] W. Yao, S. Chen, J. Cui, and Y. Lou, “Multi-stream convolutional neural network with frequency selection for robust speaker verification,” *Prepr. arXiv.2012.11159*, Dec. 2020.

BIOGRAPHIES OF AUTHORS

Pranita Niraj Palsapure is working presently as an Assistant Professor in the Department of Electronics and Communication Engineering at Acharya Institute of Technology, Bangalore, Karnataka. She is pursuing her Ph.D. under Visvesvaraya Technological University, Belgavi, Karnataka, India and M.Tech. from Nagpur University, Maharashtra in 2007. Her area of research is speech processing, machine learning. She is a member of ISTE. She can be contacted at email pranitanirajpalsapure@gmail.com.

Rajeswari is associated with Acharya Institute of Technology, Bangalore, India as Professor in the Department of Electronics and Communication Engineering. She has completed her Ph.D. in the field of speech processing. Her areas of interest include speech processing, AI, computer vision and application in the field of healthcare and agritech. She is CMI level 5 certified in management and leadership under UKIERI. She can be contacted at email rajeswari@acharya.ac.in.

Sandeep Kumar Kempegowda is presently working as Assistant Professor in Department of Electronics and Communication Engineering at Acharya Institute of Technology, Bangalore, Karnataka. He is a pursuing his Ph.D. under Visvesvaraya Technological University, Belagavi, Karnataka, India, M.E. (ECE) from Bangalore University, Karnataka in 2010. His area of research is image processing, computer vision, machine learning and embedded systems. He is a member of ISTE. He can be contacted at email sandy85gowda@gmail.com.