Journal of Information Systems and Telecommunication
Vol.12, No.4, October-December 2024, 254-263
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Enhancing Speaker Identification System Based on MFCC Feature Extraction and Gated Recurrent Unit Network

Mojtaba Sharif Noughabi [1], Seyyed Mohammad Razavi [1*], Mehran Taghipour-Gorjikolaie [1]

[1]. Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran

Received: 25 Oct 2024/ Revised: 04 Dec 2024/ Accepted: 28 Dec 2024

Abstract

One of the biometric detection methods is to identify people based on speech signals. The implementation of a speaker identification (SI) system can be done in many different ways, and recently, many researchers have been focusing on using deep neural networks. One of the types of deep neural networks is recurrent neural networks, where memory and recurrent parts are handled by layers such as LSTM or Gated Recurrent Unit (GRU). In this paper, we propose a new structure as a classifier in the speaker identification system, which significantly improves the recognition rate by combining a convolutional neural network with two layers of GRU (CNN+ GRU). MFCC coefficients that have been extracted as cell arrays from each period of Pt speech will be used as sequence vectors for the input of proposed classifier. The performance of the SI system has improved in comparison to basic methods according to experiments conducted on two databases, LibriSpeech and VoxCeleb1. When Pt is longer, the system performs better, so that on the LibriSpeech database with 251 speakers, recognition accuracy is equal to 92.94% for Pt=1s, and it rises to 99.92% for Pt=9s. The proposed CNN+GRU classifier has a low sensitivity to specific genders, which can be said to be almost zero.

Keywords: Speaker Identification; Gated Recurrent Unit Network (GRU); Convolutional Neural Network (CNN); MFCC.

1- Introduction

One of the topics of interest in various research from the past is the use of biometric features, such as face image, eyes iris, fingerprints, and voice, to recognize people. Speech biometrics can be given more attention since they don't require special equipment and can be obtained remotely through telephone lines. Voice can also aid in identifying the speaker's emotions, gender, language, and health status, in addition to conveying their identity. Our focus in this article is speaker recognition through speech signals. Speaker recognition is divided into two general subcategories: speaker identification and verification. In the identification phase after receiving the speech signal by the system, his identity is recognized, but in the verification phase, a person claims to be a specific identity using a speech signal, and the system responds to reject or validate their claim.

In these two systems, three basic stages of feature extraction, modeling, and decision-making can be used for both text-independent and text-dependent purposes [1, 2].

Mel Frequency Cepstral Coefficients (MFCC) is commonly used as a practical and important feature in experiments during the feature extraction stage. MFCC is the basis for features like MFCCT [3], SHMFCC [4], which are used in speaker recognition systems and will be explained in more detail in the next section. In addition to speaker recognition, the MFCC feature is also used in other applications such as speech emotion recognition [5]. Other features such as Power Normalized Cepstral Coefficients (PNCC) [6] and Linear Predictive Cepstral Coefficients (LPCC) [7] should also be employed during this phase. In the modeling stage, older methods such as Gaussian mixture model (GMM) and identity vector (i-vector) are used as basic methods, while Models based on deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN) are also used. Local connectivity and weight sharing in CNN reduce the number of parameters to be learned [8]. In addition to speaker recognition, the use of convolutional networks has been considered in various

✉ **Corresponding Author**
smrazavi@birjand.ac.ir

speech tasks such as speech recognition and infant cry classification [9] and image processing tasks such as person reidentification [10] and facial expression recognition [11]. Vector quantization, cosine distance, support vector machine (SVM) or neural networks are some of the methods that can be used to perform the decision-making process.

It must be pointed out that in some articles, Mel Spectrogram images or the raw speech signal are utilized for convolutional neural network input instead of feature extraction from the speech signal [12-14]. Deep neural networks are employed in three different modes in speaker recognition approaches. In the initial scenario, the network extracts features, while in the second scenario, it classifies them. In the third scenario, both feature extraction and classification are done by the deep network [15]. The second scenario has been used in this article and by employing a recurrent deep neural network, we have observed a significant improvement in system performance compared to other methods.

This paper is broken up into five parts. The subject under study was introduced in the first part, and in the second part, an overview of works relevant to the article will be provided. The method used will be explained in the third part. In the fourth part, the experiments and their results will be reviewed, and in the fifth part, the conclusions and suggestions will be presented.

2- Related Works

In this section, we will briefly review some of the research related to our work.

As previously mentioned, there are different approaches to implementing speaker recognition systems using deep neural networks, one of them was using the DNN in the classification stage. MFCC features are obtained from speech with specific lengths in [4], and a feature matrix is produced as a result. To increase the dataset, the MFCC feature vectors of every matrix are randomly arranged in terms of their placement in the matrix, without altering the vectors themselves and Form a new feature matrix together again. The name for this new feature is SHMFCC. These feature matrices are fed into a deep neural network that has five layers, consisting of one input layer, three hidden layers, and one output layer. The hidden layer is comprised of 300 neurons, a Batch Normalization layer, and a dropout layer with a probability of 0.35%. Improvement in system performance was observed during tests on two databases, LibriSpeech and VoxCeleb1.

paper [3] takes into account multiple feature vectors after extracting the MFCC feature from speeches with a certain length instead of using these vectors directly as feature vectors. By gathering 12 statistical features from these multiple vectors, a new feature vector called MFCCT was created. The new feature vector is put into a deep neural network that has 7 layers, one of which is input, five hidden layers with 200 neurons in each layer, and an output layer at the end. The proposed method has achieved relatively good results by running it on the LibriSpeech database.

Reference [16] Focuses on the use of the MFCC feature as well as other features that are commonly derived from MFCC. In the classification phase, SVM was utilized, and in the testing phase, an accuracy rate of about 90% was achieved using the ELSDSR database with only 22 speakers.

Ashar et al in [17] Achieved an accuracy of 80% for the data set with 60-speaker by extracting the 39 MFCC feature vectors from speech frames. A deep neural network with one input layer, several hidden layers, and one output layer has been used for classification. In [18], different methods are used to modify the MFCC and PNCC features, and the resulting feature vector is provided to the ELM classifier. The proposed methods for TIMIT and SITW databases achieved a maximum accuracy of 97.52 and 97.66, respectively.

Reasearchers in [19] Has achieved an accuracy of 87.65 with artificial neural networks, 89.96 with recurrent neural networks, and 99.23 with convolutional neural networks. TIMIT data has used to extract the MFCC feature of speech frames for 100 speakers. Speaker data is used to extract 12 MFCC features for each frame in [20] by considering different shapes for framing windows. The database that was utilized has 800 speeches from 16 speakers, which were prepared by the article's authors. For classification, a deep neural network with 6 hidden layers is employed. 94.37% is the average for best performance when using HANNING window.

The implementation of [21] involves the use of an open set speaker recognition system. The extracted feature is the MFCC, and the GMM-UBM model is employed during the classification process. The THYUG-20 SRE databases and speakers from noise-free parts of the LibriSpeech database were used to implement the proposed system, and accuracy levels of between 73 and 86% were achieved. In [22], an attempt has been made to enhance the speaker identification system by using reverberation modeling and techniques to cancelable speakers that can be removed. In this study, features such as wavelet-domain, MFCC and features based on DCT were extracted from the speech signal and a neural network was utilized for classification. For the experiments, the speech data of 15 Arabic speakers, including 10 men and 5 women, was utilized. The proposed method's accuracy in noisy and noiseless conditions ranges between 35 and 100%.

By extracting MFCC and MSE (Multiband Spectral Entropy) features of speech and employing various classifiers, such as KNN and DNN, the highest accuracy for ELSDSR data with 22 speakers was achieved using research [23], which resulted in 93.99% accuracy for 22 speakers. Although tests were performed on 40 speakers from the LibriSpeech database, accuracy was less than

expected. The feature vector is formed when the MFCC feature and its derivatives, along with other features like the formant frequency, are extracted and combined in [24]. This feature vector was employed in the proposed LSTM and BLSTM classifiers, which displayed a 92.75% and 95.52% accuracy rate for the YOHO database with 138 speakers, respectively.

Extracting the MFCC feature from Audio-MNIST data with 60 speakers and 500 speeches per speaker, and then using various classifiers such as SVM, KNN, LR, Nave Bayes, and so on, was done in [25]. The proposed speaker recognition system has achieved the highest accuracy of 97.1% with the SVM classifier. By extracting features from the speech signal, such as MFCC, amplitude, energy, and others, [26] was able to achieve different results, and various classification methods such as MSVM, KNN, DNN, LSTM, and Hybrid LSTM were utilized to achieve them. The Hybrid LSTM classifier achieved a high efficiency of 92.65% for 100 speakers, including 50 women and 50 men from the LibriSpeech database.

By utilizing the MFCC feature and a deep convolutional neural network for classifying, the [27] was able to achieve the highest accuracy of 94% using 251 speakers and 3-seconds long speeches. Paper [11] Has inputted raw audio signals without extracting features, simply by detecting silence in speech and separating speech parts into two different neural networks named sincNET and sincGAN. A good accuracy between 85 and 99.27% was achieved after testing these methods on TIMIT and LibriSpeech data.

The extraction of different speech features such as MFCC, PLP and PLCC and the application of classification methods such as GMM-SVM and Ivector-PLDA and their fusion using the sparse method have been done in [28]. Experiments on NIST 2004 data show better efficiency of the speaker authentication system using the sparse method. in [29] is designed a speaker recognition system by extracting the MFCC feature from speech frames and forming feature vectors from speech parts with different lengths, such as 1 second and 3 seconds and then applying various classification methods. LibriSpeech data was used to test this system and it achieved the highest accuracy of 99.31% within speeches with a length of 9 seconds. In [30], it is proposed to use Neurogram coefficients to enhance the speaker identification system's robustness. Neurogram is a 2-D time-frequency representation which was constructed by combining the neural responses (i.e., feature) from 25 AN (Auditory Nerve) fibers. The test results on the YOHO database show that the proposed method performs better than basic methods such as MFCC coefficients, especially in noisy conditions. GMM-UBM is the classification method employed.

3- Proposed Method

3-1- Feature Extraction

Our proposed methods in this article are primarily focused on classification, but some suggestions will be made for feature extraction as well. Our first task involves extracting MFCC coefficients from a speech with a specific length. Algorithm 1 is used to obtain the set of features that can be applied to the input of a recurrent deep neural network for classification purposes.

The extraction of features for each of the training, validation, and testing sections, as demonstrated in algorithm 1, results in a set of features that contain a cell array for each speech interval (Pt).

Each of these arrays is considered an input to the classifier. For example, if Pt = 1 sec, for each speech of this length, a cell array with dimensions of 13x66 is obtained. The number of MFCC coefficients in each frame is 13, and there are 66 frames in Pt's length. The number of frames is determined by the length of the frame and the amount of overlap, which can be compared with the basic methods, they are regarded as being equal to 25 and 10 milliseconds, respectively, in this article.

This cell array is inputted as a sequence into the deep neural network, which will be explained in detail later.

Algorithm1. How to Extract Features of Speaker Utterance
Input: path to speaker utterances
Segment utterances random to train, validation and test, 70, 15,15 percentage respectively
Procedure: Get MFCC Features (*path*)
M← total of class (Train or Validation or Test)
A2={} (a cell array for save total features and at end contain features cell for each class)
J← 1 (counter for classes)
While J<=M
P_t← Periods select of Utterance
N← total utterance of class J
I← 1 (counter for utterances in each classes)
A1={} (a cell array for save features that is empty each iteration)
While I <=N
A←13 MFCC features matrix from frames of utterance with P_t length
A1{I}=A
end
A2=[A2,A1]
end

It is clear that as Pt becomes larger, the number of frames and thus the length of the cell array increases. As Pt increases, the feature extraction cycle may not include certain speeches in the database because of their short length to decrease the number of speeches that are deleted,

speeches with a length of more than 0.5Pt and less than 1Pt should be continued with the part of speech that belongs to the same class until they reach the length of Pt. Of course, this part is reversed and then added to the speech. The speech is added to the feature extraction cycle after doing this.

We chose and displayed one of the speeches used in the experiments to enhance our understanding of this proposed method. As depicted in Fig. 1, the speech that was selected has a length of 11 seconds. If Pt=3, we can extract three frames with a full length of 3 seconds from this speech. However, the final frame is 1 second shorter than 3 seconds, which means it will be 2 seconds long. As this frame is longer than 0.5Pt, we'll continue with a portion of the speech that's related to the same speaker, so that its length is as long as Pt and it can be extracted from that feature. This work improves the system's performance by 3% as demonstrated by the test results. This work's results will be displayed in the test results section with the 'AUGMENTED' symbol.

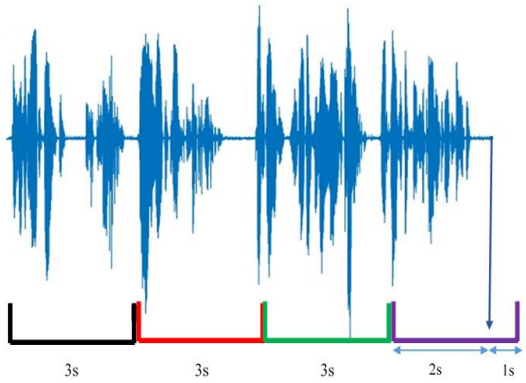

Fig. 1 How to frame speeches in experiments

The diagram in Fig. 2 illustrates the steps required to obtain the MFCC feature. Fig. 2 shows that there are eight steps to extract MFCC features from speech. These steps are explained in more detail below.

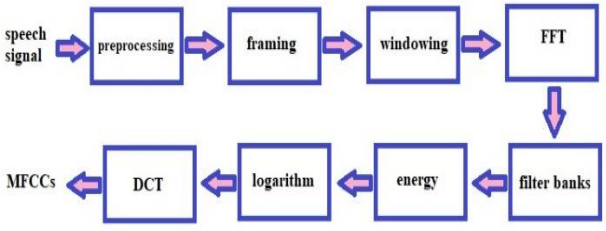

Fig. 2 Steps to calculate MFCCs

The initial step is preprocessing. In this step, a high-pass filter, also known as pre-emphasis, is applied to the speech signal to compensate for the amplitude at higher frequencies. The Eq. (1) represents this filter.

$$P(z) = 1 - az^{-1} \quad (1)$$

In the next step, the signal will be framed. The instability of the speech signal is the main reason for this action, which can be considered almost stationary because of the shortening of the speech signal in the frames. It's obvious that this action is taken to decrease the amount of input data and save time, while also analyzing the signal more closely.The frequency of the speech signal usually determines the number and length of frames. Sometimes, the signals are framed in such a way that the frames overlap with each other, and this overlap can reach up to 50%. To eliminate the discontinuity between the frames' borders, we multiply each frame in a window in the next step. The Hamming window obtained by Eq. (2) is used to calculate these coefficients.

$$w(n) = 0.54 - 0.46 * \cos\left(\frac{2k\pi}{N-1}\right) \quad k = 0,1,\ldots,N-1 \quad (2)$$

In Eq. (2), N is the length of the window, which is equal to the length of the frames. The discrete Fourier transform is performed on the windowed frames in the fourth step.

The human ear's auditory properties are the main inspiration for MFCCs. The function of the human ear is not based on physical understanding, but logarithmically and based on Eq. (3).

$$f_{mel} = 2595 \log(1 + \frac{f}{700}) \quad (3)$$

The frequency used in Equation 3 is f, while f_mel is the frequency that is converted from the linear domain to the Mel domain. The human ear's accuracy in understanding low frequencies is high, but it is low in understanding high frequencies, as shown in this equation. Mel Frequency Cepstral Coefficients are calculated using a set of filter banks to convert frequencies from Hertz scale to Mel. A triangular filter bank is the usual choice for this step. The bandwidth of triangular filters is greater at higher frequencies than at lower frequencies, which suggests that the human ear is less sensitive to frequency changes at higher frequencies than at lower frequencies. This filter bank is shown in Fig. 3.

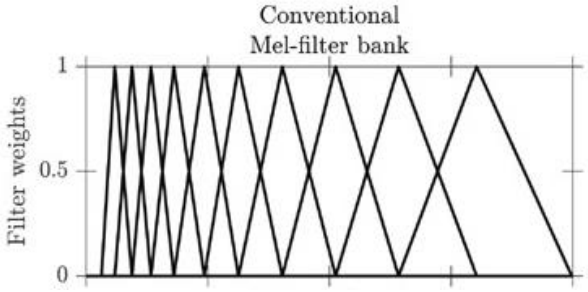

Fig. 3 Triangular filter bank [31]

After that, the energy of each of the filter banks is calculated. To decrease the numbers obtained from energy, the logarithm is employed with Eq. (4).

$$X'(m) = \log\big(X_1(m)\big) \quad (4)$$

Finally, we get the cosine transformation for the resulting coefficients by employing Eq. (5).

$$Ceps_{MFCC}(l) = \sum_{m=1}^{M} X'(m).\cos\left(l\frac{\pi}{m}.(m-\frac{1}{2})\right) \quad (5)$$

The length of each frame is M and the filter bank number is l in this equation. Mel Frequency Cepstral Coefficients are obtained by using Eq. (5) and typically yield 13 or 14 coefficients for each frame. Of course, it should be noted that in [4], in order to achieve the desired results, approximately 60 MFCC coefficients have been extracted from each frame and To improve performance in some parts of the test, non-speech parts have been eliminated before (VAD), While using our method, we have obtained better results with the same 13 MFCC coefficients without performing VAD.

3-2- Classification

In some of the articles reviewed for this step, a deep neural network with multiple hidden layers has been utilized, like in [4], where the structure of Fig. 4 is utilized in the deep neural network.

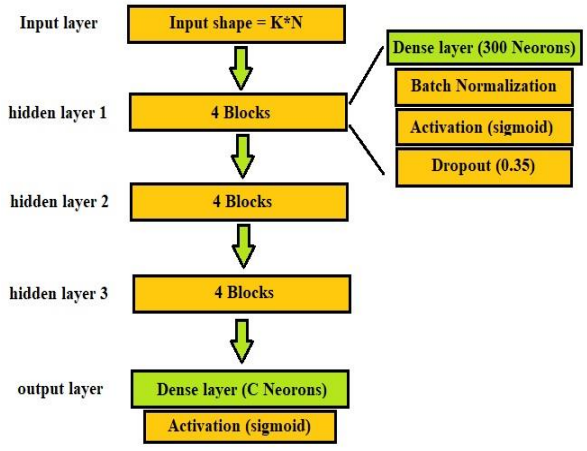

Fig. 4 The deep neural network used in [4]

The GMM-UBM model is one of the common methods used by some researches [21]. In [26], there are various methods for data classification, but the Hybrid LSTM classifier is the most efficient. The architecture of this classifier is depicted in. methods for data classification, but the Hybrid LSTM classifier is the most efficient. The architecture of this classifier is depicted in Fig. 5.

As stated in the related works section, [27] employs a CNN classifier. The proposed classification consists of 13 layers, but we choose not to display their details here. sincNET and sincGAN are utilized in [14]. These two classifications are composed of several layers, which include convolutional layers, batch normalization, and activators. For more details, refer to the reference mentioned. After performing VAD, raw audio signals are inputted into these two networks. paper [29] has presented a number of approaches for classification, including 1D-CNN, 2D-CNN, LSTM, and CRNN. There are several convolutional layers in its CNN classifier, and at the end, there is a GAP layer that enhances detection accuracy.

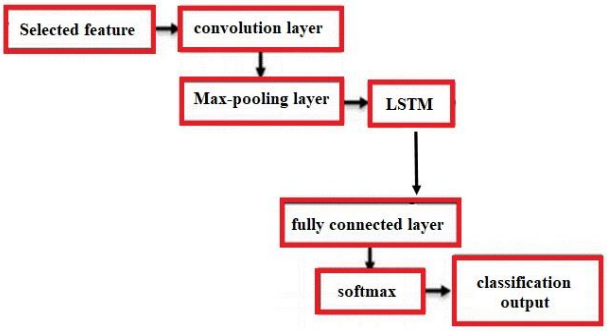

Fig. 5 Hybrid LSTM classifier architecture used in [26]

The objective of this article is to utilize a recurrent neural network that has a structure consisting of layers, as demonstrated in Fig. 5. The GRU architecture (Gated Recurrent Unit) was introduced in 2014 [32]. The purpose of this architecture is to address the shortcomings of traditional recurrent neural networks, such as gradient fading, and also to decrease the overhead of the LSTM architecture.

Deep learning models based on time series, such as Simple RNN, LSTM, and GRU, are appropriate for granting access based on previous access histories [33]. The problem of long dependency on RNN networks can be resolved with the use of GRU, a type of LSTM [34]. The module structure of GRU is repetitive and based on the attention mechanism [35], which is more straightforward than long and short-term memory because each recurrent neural network feature of the module is the same. Furthermore, unlike the LSTM with three gates, GRU has two gates: a reset gate and an update gate. The update gate is used to supervise the extent to which the knowledge of the previously hidden state is extended to the current state. The greater the value of the update gate, the more knowledge of the previous state is introduced. Therefore, if the reset gate is used to adjust the degree of knowledge transfer of the past state, the smaller the value of the reset gate, the more it will be transferred [36]. Due to its simpler structure and fewer parameters than the LSTM, the GRU neural network model can train faster and produce larger networks more easily [37].

Compared to LSTM, GRU has fewer hyperparameters and is less computationally intensive [38]. A GRU layer's internal structure is shown in Fig. 6. In this figure, Xt represents the input vector and ht represents the state

memory variable at different moments. σ is the sigmoid activation function and tanh is the tangent function.
The structure of the proposed neural network is shown in Fig. 7. This figure displays that the input of the network is sequence-based. The technique for obtaining the feature set was explained in the previous section. In this set, k is the number of MFCC coefficients, which is considered equal to 13, and N is the number of frames in the desired Pt.

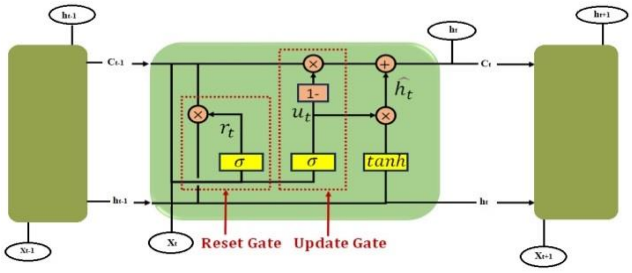

Fig. 6 A GRU layer's internal structure

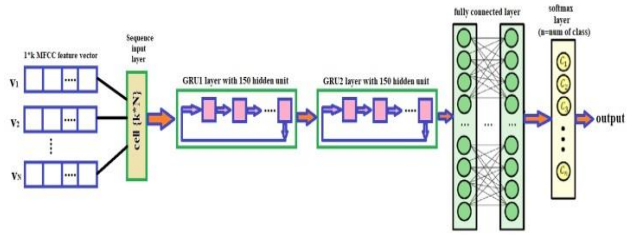

Fig. 7 structure of the proposed neural network (CNN+GRU)

First, the input speech to the system is examined and if the conditions are met, the augmented process is performed. Then the MFCC feature set is extracted from it. This sequential feature set is then fed into a GRU layer. The output of this layer is fed into another GRU layer. Each of these GRU layers has 150 hidden units. Each hidden unit has an internal state that holds information from previous inputs and uses it in the next process. The main task of the hidden unit in a recurrent network is to integrate new input information with the previous internal state. At each time t, the hidden layer receives new input information from the input layer and combines it with the previously maintained internal state.
This combination of information helps the hidden layer to recognize complex temporal patterns in sequential data. Using an LSTM layer instead of second GRU has a significant impact on the performance of the neural network, which is why we chose every double layer of GRU type.
After the GRU layers, a fully connected layer is placed to convert the features extracted from the hidden layers into an output vector. The output dimension of this layer is equal to the number of classes. finally, a Softmax activating layer is put on. The output of this layer is a probability distribution and performs the final classification task. CNN+ GRU is the name we use for this method.
Initial learning rate is set to 0.01 and the adam optimizer is employed. MATLAB 2023 software and a single GPU platform were utilized for the implementation. The speaker identification system performs better with the proposed classifier, as evidenced by the results. This classifier is not sensitive to gender, and it will be mentioned in the results section.

4- Simulation Results

In this section, we evaluate the performance of the proposed methods by evaluating them on two different databases.

4-1- Database

The experiments employed databases from the relevant articles to ensure that the results were comparable.The LibriSpeech dataset is one of the datasets, taken from the LibriVox audio book collection and has about 1000 hours of speech that are sampled at 16 KHz. The train-clean-100 set is the subset of this database that we used, and it contains speech without any noise. Of the 251 speakers in this subset, 100 speakers, including 50 men and 50 women, were selected as part of the experiment. VoxCeleb1 is another database that has been utilized. The collection contains over 100,000 speeches belonging to 1,251 celebrity speakers, taken from videos posted on YouTube. However, the speeches are not completely clean. 100 speakers from this database with an equal proportion of men and women were chosen for the experiments.
In every experiment, 70% of the data set was utilized for training, 15% for validation, and 15% for testing.

4-2- Evaluation Criteria

Choosing the appropriate evaluation criteria is essential when checking the system's performance. Speaker recognition systems can be evaluated using various criteria. In speaker recognition systems, accuracy of performance (ACC) is one of the most common criteria, and speaker recognition systems that use deep neural networks are typically evaluated with this criterion. Equal error rate (EER), MinDCF, and ROC and DET curves are used in speaker recognition and verification systems to evaluate their performance. To compare the results of the articles that have used this criterion, we use the ACC value as an evaluation criterion for the speaker identification system designed in this paper.

4-3- Results

The databases used in this paper were described in Sections 4-2. To perform the tests, the speaker's speech is segmented according to the selected Pt. The augmentation process is also performed if necessary. We divide the specified segment into 25 ms frames with 10 ms overlap and extract 13 MFCC coefficients from each frame. For each segment of speech, a feature set with dimensions 13*N is obtained, where 13 is the number of MFCC coefficients and N is the number of frames of that segment of speech. This feature set is then fed into the proposed CNN+GRU network.

The LibriSpeech database was used for our initial experiment, which involved selecting Pt values of 1, 3, and 5 seconds. Table 1 shows the test data results.

Table 1: Comparing the proposed method's ACC% results and basic methods with the LibriSpeech database

Methods	Pt (s)		
	1	**3**	**5**
MFCCT+DNN [3] (VAD , Num. class=100)	52.9	78.4	83.8
MFCC+DNN [4] (NO VAD , Num. class=100)	93.2	94.1	94.7
MFCC+DNN [23] (NO VAD , Num. class=40)	88.78	---	---
MFCC+CNN+GRU (OURS) (NO VAD , Num. class=100)	95.77	99.38	99.76
MFCC+CNN+GRU (OURS) (AUGMENTED ,NO VAD , Num. class=100)	**95.92**	**99.60**	**99.70**

Table 1 shows that the proposed method, regardless of the Pt, provides a superior output compared to the basic methods in all three cases. An improvement of more than 26% has been made when compared to method [3] and more than 4% when compared to method [4]. The results are improved by implementing the AUGMENTED method.

The VoxCeleb1 database was utilized in the next experiment. Although this set is not clean and contains background speech, the system's performance is impacted, but the proposed method still performs better. The evaluation results for the test data are shown in Table 2.

Table 2: Comparing the proposed method's ACC% results and basic methods with the VoxCeleb1 database

Methods	Pt (s)		
	1	**3**	**5**
MFCCT+DNN [3] (VAD , Num. class=100)	52.9	78.4	83.8
MFCC+DNN [4] (NO VAD , Num. class=100)	93.2	94.1	94.7
MFCC+DNN [23] (NO VAD , Num. class=40)	88.78	---	---
MFCC+CNN+GRU (OURS) (NO VAD , Num. class=100)	95.77	99.38	99.76
MFCC+CNN+GRU (OURS) (AUGMENTED ,NO VAD , Num. class=100)	**95.92**	**99.60**	**99.70**

The results of Table 2 also show that the proposed method has better performance. For two modes of Pt = 3, 5 s, there was an average improvement of more than 39% was observed in the performance of this method compared to the method [3] and more than 11% when compared to method [4]. A relative improvement in the results has been achieved by using the AUGMENTED method, just like the previous experiment.

The third experiment utilized the total LibriSpeech-clean-100 database, which has 251 speakers, with 126 male and 125 female speakers. There have been no modifications to the features of the proposed CNN+ GRU neural network, and the feature that was extracted is MFCCs. The results for the proposed method and other studied methods are shown in Table 3.

Table 3: Comparing the proposed method's ACC% results and basic methods with the LibriSpeech database

Methods	Pt (s)		
	1	**3**	**8 or 9**
Fusion of features + Hybrid LSTM [26]	92.65	---	---
MFCCT+GMM-UBM (VAD) [21]	---	---	86 (8s)
MFCC+CNN (VAD) [27]	---	94	---
RAW signals + sincNET (VAD) [14]	---	98.86	---
RAW signals + sincGAN (VAD) [14]	---	98.94	---
MFCC + 1D-CNN (VAD) [27]	90.21	97.02	99.31 (9s)
MFCC+ A-LSTM (VAD) [27]	88.48	96.98	99.22 (9s)
MFCC+ CRNN (VAD) [27]	91.98	95.94	98.10 (9s)
MFCC+CNN+GRU (NO VAD) (OURS)	**92.94**	**99.02**	**99.62** (8s) **99.92** (9s)

Table 3 illustrates that the proposed method still performs well despite the increase in speakers from 100 to 251. Regardless of the length of the speech, the proposed method has the best performance among the studied methods in all cases.

The graph in Fig. 8 is drawn to better display and compare the results obtained in Table 3.

The proposed classifier's sensitivity to the specific gender was tested in the fourth test. Identification in previous experiments was performed irrespective of the speaker's gender, which is demonstrated in the results of this section under the title of gender.

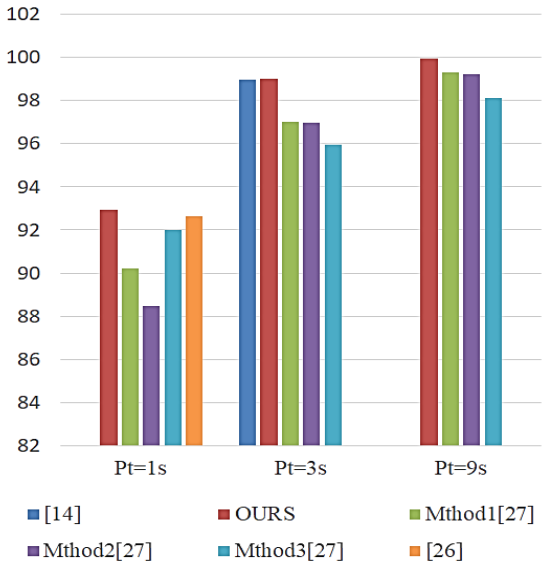

Fig. 8 A chart to compare the results of Table 3

The selected speeches from the LibriSpeech database are divided into male and female speakers in this part of the experiment and after extracting features, we insert them into the proposed classifier for identification. The results are compared in Fig. 9. The AUGMENTED method was not utilized to obtain these results.

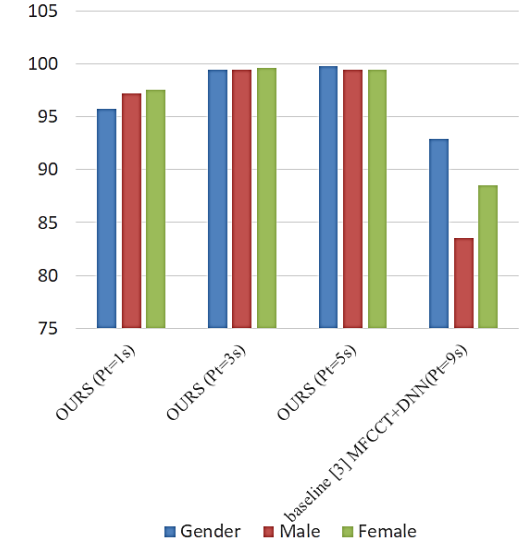

Fig. 9 The ACC% output of proposed classifier in three modes: gender, male, and female with the LibriSpeech database

Fig. 9 illustrates that the proposed classifier does not significantly decrease in performance when compared to specific genders. However, it can detect females and males better than genders in some cases.

The method used for speaker identification in this article, like many existing methods, has limitations, some of which arise during implementation. For example, training the system requires a large amount of data, and receiving long speech from speakers may not be possible in some situations. Also, in real environments, there is a possibility of noise being added to speech, which reduces the efficiency of the system. Training time also creates limitations if it is long. Of course, in this article, by using GRU layers instead of LSTM in the proposed classifier, the training time has been significantly reduced. The training time of the proposed system with different methods and the accuracy obtained are shown in Table 4. In all methods, MiniBatchSize is 52.

Table 4: Comparison of training time in the proposed system with different databases and methods

Methode: MFCC+CNN+GRU Database: LibriSpeech (100 speaker)			
Pt(s)	epochs	Traning time (min)	Acc of training (%)
1	15	28	95.84
3	28	20	99.40
5	41	21	99.19
Methode: MFCC+CNN+GRU (Augmented) Database: LibriSpeech (100 speaker)			
Pt(s)	epochs	Traning time (min)	Acc of training (%)
1	16	32	95.82
3	27	23	99.63
5	41	29	99.81
Methode: MFCC+CNN+GRU Database: LibriSpeech (251 speaker)			
Pt(s)	epochs	Traning time (min)	Acc of training (%)
1	42	104	92.91
3	60	82	99.00
8	55	39	99.72
9	70	52	99.92
Methode: MFCC+CNN+GRU Database: VoxCeleb1 (100 speakers)			
Pt(s)	epochs	Traning time (min)	Acc of training (%)
1	32	30	71.89
3	70	27	88.00
5	80	20	89.14
Methode: MFCC+CNN+GRU (Augmented) Database: VoxCeleb1 (100 speakers)			
Pt(s)	epochs	Traning time (min)	Acc of training (%)
1	50	54	72.30
3	70	35	89.14

Table 4 shows that as Pt increases, the training time decreases proportionally to the number of epochs and the accuracy of the system on the training data increases. The training time is also not long.

5- Conclusions

To enhance the performance of the speaker identification system, a convolutional neural network utilizing GRU layers was proposed in this article. Since the input to the GRU section is a sequence, the speech in the database is split into equal parts based on the considered Pt. From each part, the feature vector set of MFCCs is extracted in the form of cell arrays and sent to the proposed neural network named CNN+GRU.

The proposed method's efficiency is shown in the implementations on two different databases and with varying numbers of speakers. The system's efficiency increases as the Pt length increases. In one case, with an increase of Pt from 1s to 3s, the recognition rate increases from 71.25% to 88.87%. Increasing the length of the speech through the proposed AUGMENTED method can improve system efficiency to some extent. The proposed method also displayed a low level of sensitivity towards specific gender. It can be inferred that using the GRU layer in CNN instead of LSTM enhances both the SI system's performance and calculation speed.

References

[1] S. Hourri and J. Kharroubi, "A Novel Scoring Method Based on Distance Calculation for Similarity Measurement in Text-Independent Speaker Verification," Procedia Computer Science, vol. 148, pp. 256–265, 2019.

[2] M. Chaiani, M. Bengherabi, S. A. Selouani and M. Boudraa, "Dysarthric speaker identification with constrained training durations," 2018 International Conference on Signal, Image, Vision and their Applications (SIVA), Guelma, Algeria, 2018, pp. 1-6.

[3] R. Jahangir et al., "Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network," in IEEE Access, vol. 8, pp. 32187-32202, 2020.

[4] M. Barhoush, A. Hallawa and A. Schmeink, "Robust Automatic Speaker Identification System Using Shuffled MFCC Features," 2021 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT), Soyapango, El Salvador, 2021, pp. 1-6.

[5] S. Langari , H. Marvi, and M. Zahedi, "Efficient speech emotion recognition using modified feature extraction," Informatics in Medicine Unlocked, vol. 20, p. 100424, Jan. 2020

[6] X. Liu, M. Sahidullah and T. Kinnunen, "Optimized Power Normalized Cepstral Coefficients Towards Robust Deep Speaker Verification," 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 2021, pp. 185-190.

[7] P. Sandhya, V. Spoorthy, S. G. Koolagudi and N. V. Sobhana, "Spectral Features for Emotional Speaker Recognition," 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC), Bengaluru, India, 2020, pp. 1-6.

[8] K. Aghajani and E. P. Afrakoti I., "Speech emotion recognition using Scalogram based deep structure," International Journal of Engineering. Transactions B: Applications, vol. 33, no. 2, Feb. 2020.

[9] A. Abbaskhah, Hamed Sedighi, and Hossein Marvi, "Infant cry classification by MFCC feature extraction with MLP and CNN structures," Biomedical Signal Processing and Control, vol. 86, pp. 105261–105261, Sep. 2023.

[10] A. Sezavar, H. Farsi, and S. Mohamadzadeh, "A New Model for Person Reidentification Using Deep CNN and Autoencoders," Iranian Journal of Energy and Environment, vol. 14, no. 4, pp. 314–320, 2023.

[11] E. Ghasemi, S. M. Razavi, S. Mohamadzadeh, and M. Taghipour-Gorjikolaie, "Facial Expression Recognition through Suboptimal Filter Design Using a Metaheuristic Kidney Algorithm," Journal of Electrical and Computer Engineering Innovations, vol. 12, no. 2, pp. 425–438,2024.

[12] A.Nagrani , J. S.Chung , and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset". arXiv preprint arXiv:1706.08612. 2017.

[13] J. W. Jung , H. S. Heo , I. H.Yang , H. J. Shim , and H. J. Yu, "Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification" . extraction, vol. 8, no. 12, pp. 23-24, 2018.

[14] G. Wei, Y. Zhang, H. Min, and Y. Xu, "End-to-end speaker identification research based on multi-scale SincNet and CGAN," Neural Computing and Applications, vol. 35, no. 30, pp. 22209–22222, Aug. 2023.

[15] S. S.Tirumala and S. R. Shahamiri, "A review on deep learning approaches in speaker identification". In Proceedings of the 8th international conference on signal processing systems, Nov. 2016, pp. 142-147.

[16] K. A. Abdalmalak and A. Gallardo-Antolín, "Enhancement of a text-independent speaker verification system by using feature combination and parallel structure classifiers," Neural Computing and Applications, vol. 29, no. 3, pp. 637–651, Jul. 2016.

[17] A. Ashar, M. S. Bhatti and U. Mushtaq, "Speaker Identification Using a Hybrid CNN-MFCC Approach," 2020 International Conference on Emerging Trends in Smart Technologies (ICETST), Karachi, Pakistan, 2020, pp. 1-4.

[18] B. K. P and R. K. M, "ELM speaker identification for limited dataset using multitaper based MFCC and PNCC features with fusion score," Multimedia Tools and Applications, vol. 79, no. 39–40, pp. 28859–28883, Aug. 2020.

[19] M. K. Singh, "A text independent speaker identification system using ANN, RNN, and CNN classification technique," Multimedia Tools and Applications, vol. 83, no. 16, pp. 48105–48117, Nov. 2023.

[20] M. R. Firmansyah, R. Hidayat and A. Bejo, "Comparison of Windowing Function on Feature Extraction Using MFCC for Speaker Identification," 2021 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), Bandung, Indonesia, 2021, pp. 1-5.

[21] S. Chakraborty and R. Parekh, "An improved approach to open set text-independent speaker identification (OSTI-SI)," 2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, India, 2017, pp. 51-56.

[22] E. S. Hassan et al., "Enhancing speaker identification through reverberation modeling and cancelable techniques using ANNs," PLoS ONE, vol. 19, no. 2, p. e0294235, Feb. 2024.

[23] J. I. Ramírez-Hernández, A. Manzo-Martínez, F. Gaxiola, L. C. González-Gurrola, V. C. Álvarez-Oliva, and R. López-

Santillán, “A comparison between MFCC and MSE features for Text-Independent speaker recognition using machine learning algorithms,” in Studies in computational intelligence, 2023, pp. 123–140.

[24] N. M. Almarshady, A. A. Alashban, and Y. A. Alotaibi, “Analysis and investigation of speaker identification problems using deep learning networks and the YOHO English Speech Dataset,” Applied Sciences, vol. 13, no. 17, p. 9567, Aug. 2023.

[25] S.Hizlisoy , and , R. S. Arslan , “Text independent speaker recognition based on MFCC and machine learning”. Selcuk University Journal of Engineering Sciences, vol. 20, no. 3, pp. 73-78, 2021.

[26] V. S. R. Gade and S. Manickam, “Speaker recognition using Improved Butterfly Optimization Algorithm with hybrid Long Short Term Memory network,” Multimedia Tools and Applications, vol.13, pp.1-23, Feb. 2024.

[27] A. Fikri and A. Zahra, “Speaker Identification in Multiple Languages: Regional, Indonesian, and English with Short Utterance,” International Journal of Emerging Technology and Advanced Engineering, vol. 13, no. 9, pp. 25–35, Oct. 2023.

[28] M. Hasheminejad , and H. Farsi, (2016). “Instance Based Sparse Classifier Fusion for Speaker Verification”. Journal of Information Systems and Telecommunication (JIST), vol. 3, no. 15, pp. 1, 2016.

[29] R. Li , J. Y. Jiang , J. Liu , C. C. Hsieh , and W. Wang, “Automatic speaker recognition with limited data”. In Proceedings of the 13th International Conference on Web Search and Data Mining, Jan. 2020, pp. 340-348.

[30] Md. A. Islam, W. A. Jassim, N. S. Cheok, and M. S. A. Zilany, “A robust speaker identification system using the responses from a model of the auditory periphery,” PLoS ONE, vol. 11, no. 7, p. e0158520, Jul. 2016.

[31] S. Nagarajan, S. S. S. Nettimi, L. S. Kumar, M. K. Nath, and A. Kanhe, “Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on bark and ERB frequency scales,” Digital Signal Processing, vol. 104, p. 102763, Sep. 2020.

[32] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” arXiv (Cornell University), Jan. 2014.

[33] N. Mohammadi, A. Rezakhani, H. H. S. Javadi, and P. Asghari, “FLHB-AC: Federated Learning History-Based Access Control using Deep Neural Networks in healthcare system,” Journal of Information Systems and Telecommunication (JIST), vol. 12, no. 46, pp. 90–104, Jun. 2024.

[34] A. Shewalkar, D. Nyavanandi, and S. A. Ludwig, “Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU,” Journal of Artificial Intelligence and Soft Computing Research, vol. 9, no. 4, pp. 235–245, Oct. 2019.

[35] A. Barati, H. Farsi, and S. Mohamadzadeh, “Integration of the latent variable knowledge into deep image captioning with Bayesian modeling,” IET Image Processing, , vol. 17, no. 7, pp. 2256–2271,2024.

[36] H. S. Munir, S. Ren, M. Mustafa, C. N. Siddique, and S. Qayyum, “Attention based GRU-LSTM for software defect prediction,” PLoS ONE, vol. 16, no. 3, p. e0247444, Mar. 2021.

[37] C. Yin, D. Tang, F. Zhang, Q. Tang, Y. Feng, and Z. He, “Students learning performance prediction based on feature extraction algorithm and attention-based bidirectional gated recurrent unit network,” PLoS ONE, vol. 18, no. 10, p. e0286156, Oct. 2023.

[38] Y. Wang et al., “Prediction of outpatients with conjunctivitis in Xinjiang based on LSTM and GRU models,” PLoS ONE, vol. 18, no. 9, p. e0290541, Sep. 2023.