Speaker identification under noisy conditions using hybrid convolutional neural network and gated recurrent unit

Wondimu Lambamo[1], Ramasamy Srinivasagan[2], Worku Jifara[1], Ali Alzahrani[2]
[1]Department of Computer Science and Engineering, School of Electrical Engineering and Computing,
Adama Science and Technology University, Adama, Ethiopia
[2]Department of Computer Engineering, King Faisal University, Al Hofuf, Saudi Arabia

Article Info

Article history:

Received Jul 15, 2023
Revised Sep 3, 2023
Accepted Oct 5, 2023

Keywords:

Cochleogram
Convolutional neural network
Gated recurrent unit
Signal to noise ratio
Speaker identification

ABSTRACT

Speaker identification is biometrics that classifies or identifies a person from other speakers based on speech characteristics. Recently, deep learning models outperformed conventional machine learning models in speaker identification. Spectrograms of the speech have been used as input in deep learning-based speaker identification using clean speech. However, the performance of speaker identification systems gets degraded under noisy conditions. Cochleograms have shown better results than spectrograms in deep learning-based speaker recognition under noisy and mismatched conditions. Moreover, hybrid convolutional neural network (CNN) and recurrent neural network (RNN) variants have shown better performance than CNN or RNN variants in recent studies. However, there is no attempt conducted to use a hybrid CNN and enhanced RNN variants in speaker identification using cochleogram input to enhance the performance under noisy and mismatched conditions. In this study, a speaker identification using hybrid CNN and the gated recurrent unit (GRU) is proposed for noisy conditions using cochleogram input. VoxCeleb1 audio dataset with real-world noises, white Gaussian noises (WGN) and without additive noises were employed for experiments. The experiment results and the comparison with existing works show that the proposed model performs better than other models in this study and existing works.

Corresponding Author:

Wondimu Lambamo
Department of Computer Science and Engineering, School of Electrical Engineering and Computing
Adama Science and Technology University
Adama, Ethiopia
Email: wondimuwcu@gmail.com

1. INTRODUCTION

Speaker recognition [1] is a biometric mechanism that classifies or identifies individuals from other known speakers by using speech characteristics. The speech of two different people cannot be the same because of differences in physiological and behavioral features of the speaker in speech [2]. The structure, size, and shape of speech production organs of each person are unique. Moreover, every person has different styles of speaking, timing of words, word selection, and so on [3]. Speech-based systems make the interactions between humans and machines natural. The accessibility of speech, acceptability by users, and availability of low-cost resources for implementation have increased the need for speech-based systems in various areas. Speaker recognition has attracted the interest of researchers and users because of its advantages in various real-world applications. Speaker recognition technologies can be applied from small business areas to large organizations for various purposes. For example, it is important in speech-based authentication systems, security [4], forensics investigation [5], surveillance [6], automatic identity tagging, and front-ends of automatic speech recognition [7].

Speaker recognition can be classified into identification and verification, depending on how the system classifies the speaker based on the claimed speech [8]. In speaker verification, the system determines a person giving the utterances to the system whether exists or not. To make a decision, the system computes the probability of similarity between classes. Then the system compares the probability with the threshold values to decide whether to authorize or reject. Speaker verification uses binary classification models [9] in which the speaker is either authorized or rejected based on the threshold of similarity. This is an open-set classification in which the test samples are assumed to be either trained or untrained. In this study, we focused on developing a speaker identification model. Speaker identification is a type of biometric that automatically determines the person who gives the speech to the system from a set of trained speakers [10]. In speaker identification, the model computes the probability of test speech similarity with each of the speaker classes trained by the model. The class of the speaker with the maximum similarity can be determined as the speaker of utterance. This is sometimes referred to as closed-set speaker recognition, in which all the test speeches are assumed to be from a set of registered speakers. Feature extraction and classification models are two basic components of speaker identification that play a crucial role in the performance of the models [11]. Conventional machine learning methods [12] and deep learning models [13] have been widely employed for speaker identification. Our study is focused on developing a speaker identification model using deep neural network models.

The speaker recognition system performance gets degraded because of environmental noises, channel variations, physical changes in the speaker, language variations, and behavioral changes in the speaker. These factors are the main challenges for implementing speaker identification systems in various real-world applications. Several studies have been conducted using both conventional machine learning and deep learning methods to enhance the accuracy of speaker recognition systems under noisy and mismatched conditions.

Some of the conventional machine learning methods include the Gaussian mixture model (GMM) [14] and support vector machine (SVM) [15]. These methods have been using handcrafted features for speaker recognition and other speech analysis purposes. Mel frequency cepstral coefficient (MFCC) and gammatone frequency cepstral coefficient (GFCC) are some of the commonly employed handcrafted features. In the studies [16] and [17], MFCC features were employed in speaker verification systems using i-vector and GMM classifiers respectively. The study [18], employed MFCC features and GMM to develop a speaker recognition system. In the studies [19], [20] speaker recognition systems using the GMM and MFCC features were proposed as the biometrics for smart home device control and remote identification over the voice-over-internet protocol (VoIP) respectively. Speaker recognition systems using MFCC features performed well on clean speech and without mismatched training and test speech. However, the accuracy of speaker recognition systems using MFCC features gets degraded with real-world noises, background noises, and changes in the physical and behavioral characteristics of speakers [21]. In the research works [22], [23] GFCC features surpassed the accuracy of MFCC features in speaker recognition under environmental noises. Therefore, GFCC features have been widely employed for speaker recognition under noisy and mismatched conditions to enhance the accuracy of systems. For instance, in the study [24] GMM method and GFCC features were employed for speaker identification under noisy acoustic datasets. Another study in [25], proposed a speaker identification system for forensic applications using the GMM with universal background model (GMM-UBM) and GFCC features in noisy environments. In the reference [26] a GMM method and GFCC feature were employed for speaker verification using datasets with real-world noises.

Recently, deep learning models outperformed conventional machine learning methods in the area of speech analysis including speaker recognition. Convolutional neural network (CNN) and recurrent neural network (RNN) [27], are some of the commonly employed deep learning models in speaker identifications. CNN model extracts feature automatically from the input and learn adaptively from the parameters. It has advantages in extracting short-term features from input data and extracts a limited number of training parameters for model training which require low computational resources [28]. Enhanced variants of the RNN model include gated recurrent unit (GRU) [29], long short-term memory (LSTM), and bidirectional LSTM (BiLSTM) [30]. The GRU model is an enhanced version of the LSTM and it has the advantages of extracting and learning long-term correlation between features sequentially in only one direction [31]. GRU models have fewer parameters than LSTM to minimize the computation cost. Moreover, GRU models handle gradient vanishing problems and converge within a few iterations during model training. In recent studies, a hybrid network of CNN models with enhanced variants of RNN models has been showing better performance in various areas. Most of the recent speaker identification methods employed either CNN or variants of RNN models.

Some deep learning-based speaker identification models employed hand-crafted features which were common in conventional machine learning methods. In the study [32], [33] speaker identification models were developed using CNN and MFCC features. Another study in [34], employed the MFCC features for speaker identification using the visual geometry group (VGG) version of the CNN model. The speaker verification model using a Siamese network of CNN and MFCC features was proposed for cross-device platforms [35]. GFCC features were used together with the CNN model to develop a speaker identification model under noisy conditions [36]. Although MFCC and GFCC features have comparatively better performance in deep learning than conventional machine learning, they are not as efficient as other inputs in deep learning for speaker recognition. Spectrograms of the speech were widely employed in deep learning models for speaker recognition. In the study [37], spectrogram features showed superior performance than MFCC features in speaker recognition using a CNN model. In the study [38], speaker recognition using spectrogram features performed better than the raw waveforms of the speech and MFCC input in deep learning. This is because spectrogram features are rich in acoustic features of the speaker which helps the deep learning networks easily learn the correlation between the features. However, the performance of speaker identification using a deep learning model and spectrogram of speech input gets degraded under noisy and mismatched conditions. Cochleograms of the speech have achieved better accuracy than spectrogram features in speaker identification under noisy and mismatched conditions. In the study [39], we analyzed the noise robustness of cochleogram and spectrogram features in speaker recognition using different kinds of CNN architectures. The analysis results show that cochleogram features have better performance than spectrogram features under noisy conditions and both features have approximately similar performance in clean environments.

In recent studies, a combination of CNN with variants of RNN has shown better performance in various areas. In the studies [40], [41] a hybrid network of CNN and LSTM models exhibited performance improvement in speaker verification and identification. Another study [42], employed a hybrid network of CNN and BiLSTM for language identification using spectrograms of speech and the results have shown improvement in existing works. The study in [43], also indicated the effectiveness of a hybrid network of CNN and BiLSTM models in audio-visual recognition for biometric applications. However, there is no attempt conducted by using hybrid CNN and GRU models which used cochleogram of speech input to enhance the accuracy of speaker identification under noisy conditions.

As discussed above, most of the research works in speaker recognition were conducted by using deep learning models CNN or RNN variants and spectrograms of speech input. Only a few studies employed a hybrid network of CNN and RNN variants for speaker recognition. However, there is no attempt conducted to use hybrid CNN and variants of RNN in speaker identification under noisy and mismatched conditions. In addition, none of the speaker recognition models which employed a hybrid deep learning model used cochleograms as input to enhance accuracy under noisy conditions. In this study, a speaker identification using a hybrid CNN and GRU model on the cochleograms of the speech input was proposed for a noisy condition. The proposed model integrates the advantages of the CNN and GRU layers in feature extraction, learning and classification. Cochleogram features were used as the input because of the rich acoustics features of the speaker and it can represent low-frequency samples in high resolution which is important in noisy speech. CNN models have the advantage of automatically extracting short-term feature dependencies and adaptive learning from the parameters. The GRU models have advantages for extracting long-term one-directions correlation of features sequentially, handles gradient vanishing and exploding problems, and converge faster during the experiment. Experiments were conducted using the VoxCeleb1 audio dataset with real-world noises (babble, restaurant and street), white Gaussian noise (WGN) and without additive noises. The real-world and WGN were added to each utterance at the signal to noise ratio (SNR) of -5 to 20 dB in the interval of 5 dB. Additional experiments were conducted by using two-dimensional CNN (2DCNN), hybrid CNN with LSTM (CNN-LSTM), and hybrid CNN with the BiLSTM (CNN-BiLSTM) to present the effectiveness of the proposed model. Comparisons with the existing works were also conducted to show the effectiveness of the proposed model.

The rest of the paper is organized as follows: Section 2 presents the theoretical background of the study which includes convolutional neural network, gated recurrent unit and cochleogram generation processes. Section 3, discusses the methodology. Section 4, presents results and discussion which include datasets, implementation details and results. In section 5, the conclusion and future works are presented.

2. THEORETICAL BACKGROUND

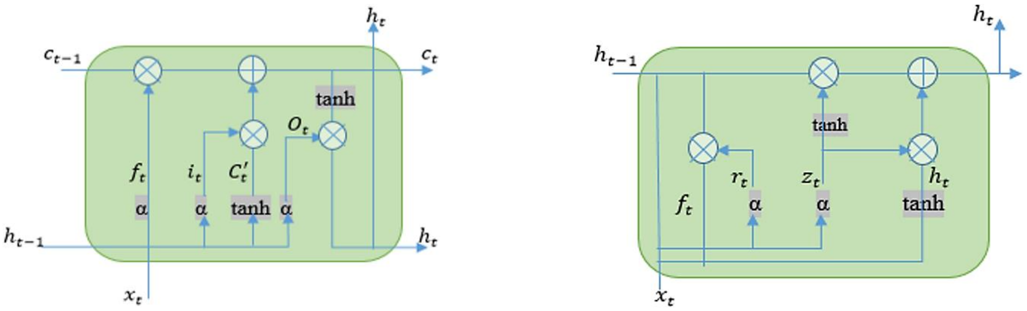
2.1. Convolutional neural network (CNN)

A convolutional neural network (CNN) is a feed-forward network of deep learning layers that has no cycle and memory. CNN models automatically extract features from the input without human intervention. Various types of CNN models were applied in areas such as medical image classification, computer vision, biometric recognition, speech recognition, and so on. CNN models represent each class with few parameters, which saves computational resources and it has a strong learning ability from few parameters. It comprises different types of layers that are interconnected to each other for feature extraction, learning the pattern, and classifying. The basic layers in CNN models include the convolutional layer, pooling layer, fully connected, and classification layers. A convolution layer is the most important component for building CNN models.

Important parameters of convolution layers are filters, activation, padding and so on. Filters extract parameters or features from the input image and each filter parameter is important to train the model throughout the training. Filter size should always be smaller than the size of the input image. Each filter convolves with the image and creates an activation map. For convolution, the filters move across the height and width of the image and the dot product between every element of the filter and the input is calculated at every spatial position. Pooling layers are commonly applied for the subsampling of feature maps. It should be applied following the convolution layers to minimize the dimension of the features for the next layers. There are several pooling methods such as tree, gated, average, min, max, global average and so on. Max, min, and global average pooling were widely employed in most deep learning-based applications.

2.2. Gated recurrent unit (GRU)

Recurrent neural network (RNN) is one of the deep learning models that has been widely employed in time series data analysis. It has a memory to store information about the previous sequences and forward the important state of the previous state to the next layers based on the gate's decision. RNN models extract the correlation between the features sequentially in a time series. However, standard RNN models can be affected by gradient exploding and vanishing in the features that have a long sequence of dependency. Standard RNN may fail to capture the long-term temporal correlation between the features because of its gradient vanishing and gradient exploding problem. Enhanced variants of RNN have been employed to solve the problems of standard RNN by storing long-term temporal dependency between the features. LSTM and GRU are widely applied variants of RNN in speaker recognition and other areas. LSTM networks have three types of gates such as input, forget, and output gates. In the input gate, the information that should be stored in the long-term memory is decided. The information that should be discarded from the long-term memory and that should be kept in the long-term memory is decided by the forget gate. New short-term memory information gets generated from the current input, previous short-term memory information and newly generated long-term memory on the output gate. LSTM models are more complex because they consist of more gates and a larger number of parameters than GRU. An overfitting problem and high computational resource consumption are the disadvantages of LSTM networks. GRU is one of the variants of RNN which uses gates to regulate, manage and manipulate information in the cells of the neural network. GRU models have only two gates (i.e., Update and reset gates) that can handle the limitation of LSTM models. In GRU, the update gate shows the combinations of the input gate and forget gate of LSTM models. The update gate of GRU models mainly controls the amount of information that should pass from the previous state to the next state. The update gate minimizes the risk of gradient vanishing and exploding problems by storing the information of the previous state of long sequence dependency. In the reset gate of the GRU models, the amount of previous information that should be discarded is decided (i.e., decides whether the previous cell state is useful or not). GRU architectures are simple in structure because it has only two types of gates in each of the cell. The cell structures of the RNN variants are shown in Figure 1 [44]. Figure 1(a) shows the cell structure of LSTM networks and Figure 1(b) shows the cell structure of the GRU networks.



(a) (b)

Figure 1. The cell structure of RNN variants, (a) the cell structure of LSTM and (b) cell structure of GRU

2.3. Cochleogram generation

Cochleogram is the representation of speech in a two-dimensional (2D) time-frequency image to employ in speech analysis tasks. The x-axis of the cochleogram represents time, the y-axis represents frequency, and the color represents the amplitude of the speech sample in the image. The RGB (i.e., red, green, blue) or grayscale are commonly useful colors to represent the amplitude of the speech during speech analysis.

The cochleogram generation process comprises pre-emphasis, framing, windowing, fast Fourier transform, and gammatone filter bank as shown in Figure 2.

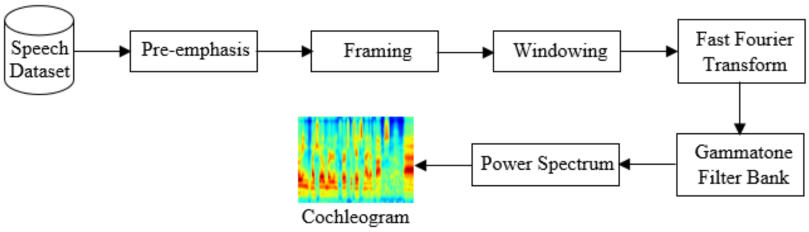


Figure 2. Cochleogram generation process

Pre-emphasis compensates the energy in the high-frequency sample concerning energy in the low-frequency samples. For the speech sample s[n] and emphasis factor p, the pre-emphasized speech sample y[n] can be computed according to (1).

$$y[n] = s[n] - p * s[n-1]; \qquad 0.91 \leq p < 1 \tag{1}$$

The speech signal is very dynamic and obtaining stable information from the long speech sample is difficult. Framing the speech into fixed short samples is important to find stable acoustic information from the speech samples. The recommended frame duration to find stable acoustic information ranges from 20ms to 30ms, then each adjacent frame should overlap from 30% to 50%. During framing the edges of the segments become sharp, the mismatch between the original sample and the segment may occur, and discontinuity between the segments also happens. Windowing reduces the effect of framing on the sample, which handles unexpected changes, undesirable frequencies and smoothes the edges of segments. Hamming window have been widely employed in speech analysis. Each frame Y[n] can be windowed as shown in (2), and the window function W[n] is calculated as shown in (3).

$$X[n] = Y[n] * W[n] \tag{2}$$

$$w[n] = 0.5 - 0.46\, cos\left(\frac{2n\pi}{M-1}\right), 0 \leq n \leq M-1 \tag{3}$$

In the time domain of the speech signal, obtaining the acoustic characteristics of the speaker is difficult. The speech signal in the frequency domain is better for finding the acoustic characteristics of the speaker. Fast Fourier transform (FFT) computes the equivalent frequency domain of the samples of the time domain, the result is referred spectrum or periodogram. FFT of the input frame X (n) and N number of points can be computed according to (4).

$$\mathrm{X(k)} = \sum_{\mathrm{n=0}}^{\mathrm{N-1}} \mathrm{X(n)e}^{-\frac{2\mathrm{j}\pi\mathrm{nk}}{\mathrm{N}}}; \qquad 0 \leq \mathrm{k} \leq \mathrm{N} - 1 \tag{4}$$

Cochleogram can be obtained from the gammatone filters in the equal rectangular bandwidth (ERB) scale. ERB measures the psychoacoustics of the speech to determine the approximate bandwidths of the filters in human hearing. Gammatone filters simulate the human auditory system which helps to model the speaker, language, speech and other information from the speech. It has advantages in representing the speech of low frequency in finer resolution. The cochleogram generation process follows similar steps with that of the spectrogram generation, except gammatone filters are used instead of Mel filters. Both GFCC and Cochleogram features can be generated from the gammatone filter which makes them preferable for speech analysis under noisy conditions. For the FFT of the speech frames with amplitude *a* and time t, gammatone filter of order n and phase shift φ can be computed according to (5). The central frequency f_{cm} of the m^{th} gammatone filter computed from low-frequency f_L and high-frequency f_H according to (6). The ERB scale of the fcm was calculated as in (7). The bandwidth b_m of the gammatone filters can be calculated as shown in (8).

$$g(t) = at^{n-1}e^{-2\pi b_m t}\, cos(\pi f_{cm} t + \varphi) \tag{5}$$

$$f_{cm} = \left(-\frac{1000}{4.37}\right) + \left(f_H + \frac{1000}{4.37}\right) exp\ (\frac{m}{M}(-\, ln\left(f_H + \frac{1000}{4.37}\right) + ln\ (f_L + \frac{1000}{4.37})) \tag{6}$$

$$\mathrm{ERB}(\mathrm{f_{cm}}) = 24.7\left(4.37 * \left(\frac{\mathrm{f_{cm}}}{1000}\right) + 1\right) \tag{7}$$

$$b_m = 1.019 * ERB(f_{cm}) \tag{8}$$

Each gammatone filter can be used as a feature to model the speaker, speech, language, emotion and other information by using the speech. Gammatone filters of the given speech in the ERB scale are stacked together to generate the cochleogram of the given speech. In this study, cochleogram have been used to model the speaker characteristics and noise information from the speech. Sample speech waveform and its cochleogram are presented in Figure 3. The speech waveform is shown in Figure 3(a) and cochleogram is presented in Figure 3(b).

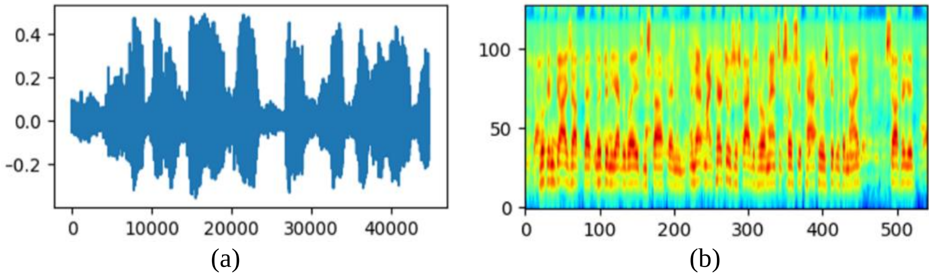


Figure 3. Speech waveform and its cochleogram, (a) speech waveform and (b) cochleogram

3. METHODOLOGY

The proposed model consists of three basic components such as Cochleogram generation, CNN layers and GRU layers. Cochleogram generation module has been used to convert each of the speech in the dataset into the cochleogram. CNN layers were employed to automatically extract short-term correlations between the features in the cochleogram input. GRU layers were employed to extract long-term feature dependency sequentially in one direction. The proposed model architecture is illustrated in Figure 4.

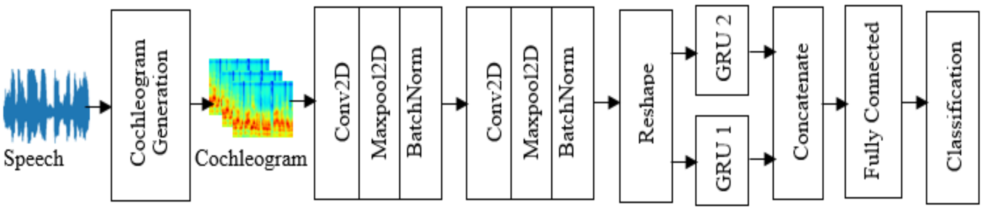


Figure 4. Proposed model architecture

Cochleogram of the speech has been used as an input for training and testing the proposed model. Therefore, each of the utterances in the dataset was converted into cochleogram to feed the input tensor of the model. Cochleogram generation was performed according to the sequence of operations in Figure 1. First, speech is pre-emphasized with a factor of 0.97, then framed into segments of 30 ms by overlapping adjacent frames with 10ms and each frame passes through the hamming window. FFT of 128 filters and 2,048 filter points were extracted from each of the windowed frames. To compute gammatone filters from each FFT, the lower frequency was set to zero Hz and the higher frequency was set to the half of frame’s sample frequency which is 8,000 Hz. The central frequency of the gammatone filter was computed from the lower and higher frequencies according to (6) and its ERB scale was computed based on (7). The bandwidth of the gammatone filters was computed from the ERB scale of central frequency according to (8). Gammatone filters were computed from each of the fast Fourier transform using central frequency, bandwidth, time, amplitude, and filter order as input according to (5). Each of the gammatone filters power spectrum was stacked together to generate cochleogram of the given speech. The cochleogram of size 160x160x3 was used as an input to the proposed model.

The proposed model consists of two convolution layers which are followed by two GRU layers. Two of the convolution layers were employed consecutively at the beginning to extract short-term spatial feature correlation. The kernel size of 3x3, the same padding, ReLu activation and normal kernel initializer were employed in each of the convolution layers. The size of filters in the first and second convolution layers were 16 and 32, respectively. To reduce the dimension of features, the maxpooling layer of size 2x2 has been employed after each of the convolution layers. Batch normalization was also employed after each maximum pooling layer to normalize the features.

The GRU layers were employed after the last maxpooling layer. The output shape of the maxpooling layer is not similar to the input shape of the GRU layers. A reshape layer was connected between the maxpooling and GRU layers to convert the output shape of the maximum pooling layer into an input shape of the GRU layer. The output of the reshape layer was provided as an input for both GRU layers. The number of cell units in each GRU layer is 256 units. A normal kernel initializer was used in each GRU layer. A concatenate layer was employed to concatenate the output of the GRU layers. A single fully connected layer was used after the concatenation layer to train the networks of the proposed model. The fully connected layer used the output of the concatenate layer as the input. The softmax activation function was used to identify and classify the speaker. The number of classes used in the softmax function was 1,251 which is equal to the total number of speakers in the dataset. Implementation detail of the proposed model is presented in Table 1.

Table 1. Implementation detail of proposed model

Layer No.	Layer name	Output shape	Param #
1	Input	(None, 160, 160, 3)	0
2	Conv2D	(None, 160, 160, 16)	448
3	MaxPooling2D	(None, 80, 80, 16)	0
4	Conv2D	(None, 80, 80, 32)	4640
5	MaxPooling2D	(None, 40, 40, 32)	0
6	Reshape	(None, 40, 1280)	0
7	GRU	(None, 256)	1181184
8	GRU	(None, 256)	1181184
9	Concatenate	(None, 512)	0
10	Dense	(None, 1251)	641763
11	Softmax	(None, 1251)	0
Total params: 3,009,219			
Trainable params: 3,009,219			

4. RESULTS AND DISCUSSION

4.1. Dataset

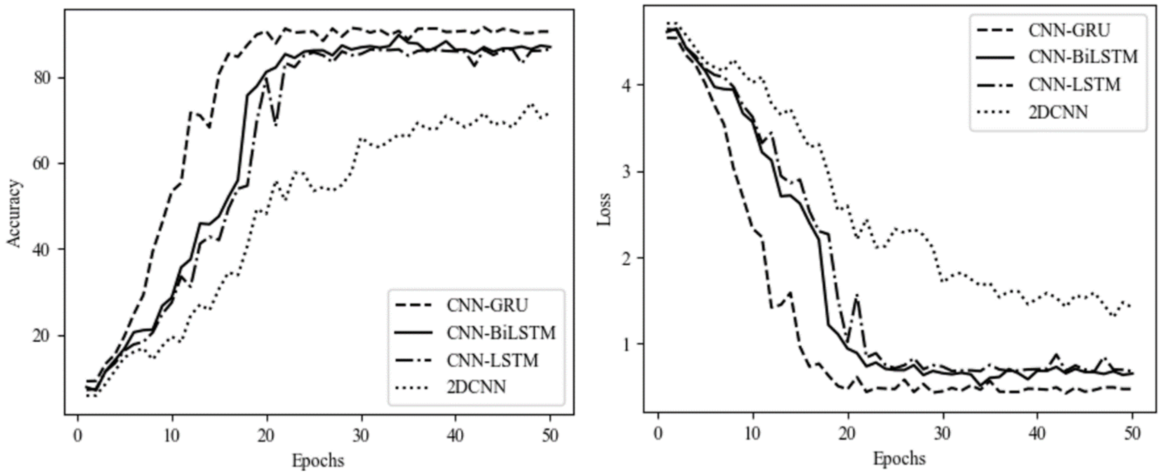
The performance of the proposed model and other models in this study were evaluated by using the VoxCeleb1 audio dataset. VoxCeleb1 audio dataset is an open-source dataset, primarily collected for speaker recognition tasks. It consists of 153,516 utterances that were collected from 1,251 speakers. All utterances in the dataset were extracted from various celebrity videos uploaded on YouTube. The proportion of male speakers in the dataset is approximately 55%. The utterances of the dataset were collected from videos of English language speakers with different types of accents. The dataset contains utterances with various types of accents, ethnicities, gender, profession, age, and speaking style. Utterances contain environmental and channel noises. Each utterance was collected at a sample rate of 16,000 Hz. The number of utterances in each speaker class is not equal (e.g., the minimum and maximum number of utterances in the VoxCeleb1 dataset classes are 45 and 1,002 respectively). The original training and test split of the dataset was first mixed into one. Then, we split the dataset into training, validation and test sets with the ratio of 80%, 10%, and 10%, respectively. The number of utterances in the splits was not equal for each speaker class because each class had a different number of utterances. Although the VoxCeleb1 dataset contains noises in various ratios, it was assumed as a clean dataset. The models are evaluated by using the dataset with real-world noises, WGN and without additive noises. The VoxCeleb1 dataset with the real-world noise was generated by adding one randomly selected real-world noise (i.e., babble, restaurant and street) to each utterance in the dataset. The VoxCeleb1 dataset with the WGN was generated by adding WGN which is generated using Python code to each utterance at different noise ratios. Real-world noises and WGN were added to each utterance at the SNR of -5 dB to 20 dB in the interval of 5 dB.

4.2. Implementation details

The experiments in our study were conducted using the TensorFlow library developed for machine learning and deep learning in Python. The Anaconda navigator was employed to manage important packages and was used to deploy the experiment. JupyterLab 3.5.3 was used to write and run code using Python language. Speech analysis to generate cochleogram was conducted using Spafe (simplified python audio features extraction) package. We conducted experiments using NVIDIA TITAN Xp graphics processing unit (GPU). The NVIDIA TITAN Xp GPU has better efficiency than central processing unit (CPU) based computers for deep learning tasks. The initial batch size and learning rate of the model were set as 32 and 0.0001, respectively. The optimization function of the root mean squared propagation (RMSprop) optimizer was used to obtain the optimum results of the model. The categorical cross-entropy loss function was applied to compute the loss of the model. The models were trained for 50 epochs, at each of the epochs the model's training and validation results were computed based on the metrics.

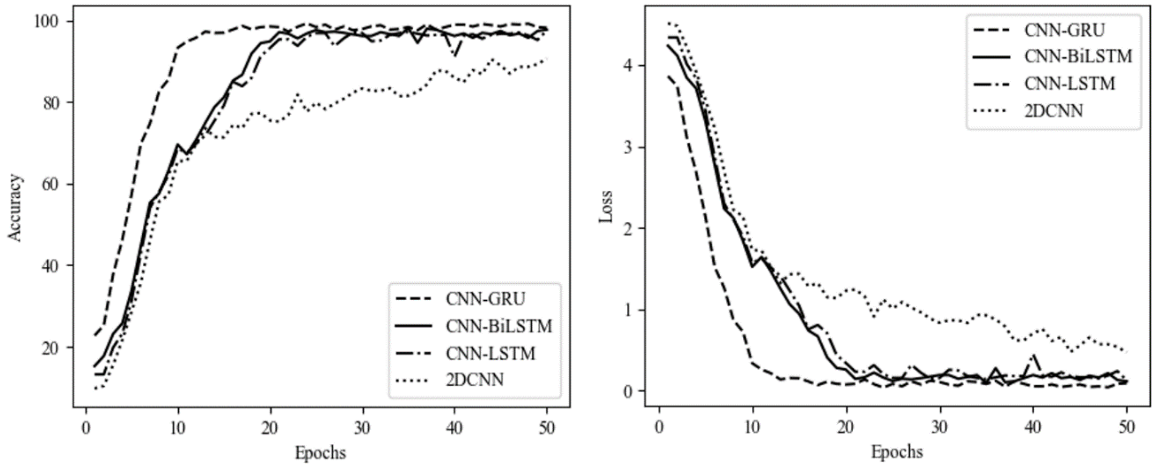
4.3. Results

The experiment of each model was repeated for 10 rounds and the average of the results was used to report the results of each model. To illustrate the performance of the models experimented in this study, one training progress of each model under high and low real-world noises were selected. The performance of each model under high real-world noise (i.e.; at SNR=-5 dB) is presented in Figure 5. The Figure 5(a) presents the accuracy of models and Figure 5(b) shows the loss of models. The performance of the models on the dataset with low real-world noise ratio (i.e.; at SNR=20 dB) is presented in Figure 6. The Figure 6(a) and Figure 6(b) presents the accuracy and loss of models at low real-world noise respectively.



(a) (b)

Figure 5. Performance of models on the dataset with high real-world noise ratio (i.e.; SNR=-5 dB), (a) accuracy of models at high real-world noises and (b) loss of models at high real-world noises



(a) (b)

Figure 6. Performance of models on the dataset with low real-world noise ratio (i.e.; SNR=20 dB), (a) accuracy of models at low real-world noises and (b) loss of models at low real-world noises

Table 2, presents an average accuracy of different types of deep learning models in speaker identification with the WGN at the SNR range from -5 dB to 20 dB. The average accuracy of the models at SNR equals -5 dB and 0 dB ranges from 71.66% to 91.82% and 83.7% to 94.22% respectively. The average accuracy of the models at SNR equals 5 dB and 10 dB ranges from 86.61% to 95.96% and 87.48% to 97.52% respectively. At SNR equals 15 dB and 20 dB the average accuracy of the models ranges from 88.90% to 98.05% and 90.78% to 98.43% respectively. The results show that 2DCNN has the lowest accuracy than other models at different ranges of SNR, it has an average accuracy range from 71.66% to 90.78%. CNN-GRU model have the highest accuracy than other models at all the SNR level, it has an average accuracy which ranges from 91.82% to 98.43%. At SNR equals -5 dB, the CNN-GRU model has shown improvements of 4.24% to 20% on other models. At SNR equals 20 dB the CNN-GRU model achieved an improvement of 0.9% to 7.65% accuracy. Generally, the proposed CNN-GRU model has better performance under noisy conditions at a different level of SNR.

Table 2. Speaker identification performance of deep learning models using VoxCeleb1 dataset with WGN

Method	Accuracy (%) at SNR					
	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
2DCNN	71.66	83.7	86.61	87.48	88.90	90.78
CNN-LSTM	86.61	89.85	93.86	96.35	97.07	97.36
CNN-BiLSTM	87.58	91.51	94.72	96.44	97.18	97.55
CNN-GRU (Proposed)	91.82	94.22	95.96	97.52	98.05	98.43

Figure 7, presents a comparison of speaker identification performance of different types of deep learning models experimented on in this study on the WGN-added VoxCeleb1 dataset at SNR ranges from -5 dB to 20 dB. The figure shows that the CNN-GRU model has the highest performance in all SNR levels. 2DCNN has the least performance in all the SNR levels. Both CNN-LSTM and CNN-BiLSTM have relatively similar performance. In speeches with high noise, the CNN-GRU model has shown the highest improvement on the other models.

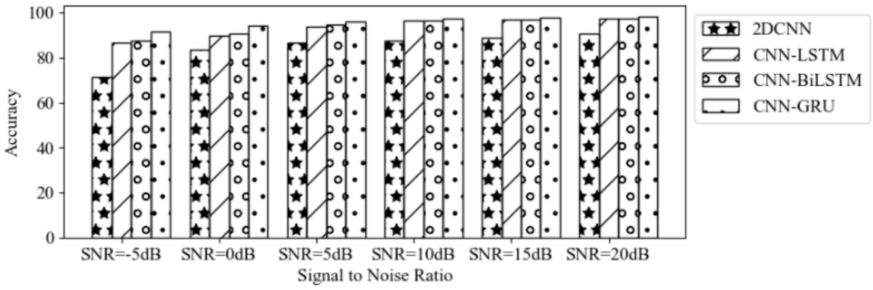


Figure 7. Speaker identification performance of models using dataset with WGN

Table 3, reports an average speaker identification accuracy of deep learning models employed in this study on the real-world noise added VoxCeleb1 dataset at the SNR from -5 to 20 dB. The average accuracy of the models on the dataset with high real-world noise (at the SNR equals -5 dB and 0 dB) ranges from 71.48% to 91.33% and 83.52% to 94.03% respectively. On the dataset with medium real-world noise (at SNR equals 5 dB and 10 dB), the accuracy of the models ranges from 86.42% to 95.76% and 87.28% and 97.31% respectively. The accuracy of the models on the dataset with low real-world noise (at SNR equals 15 dB and 20 dB) ranges from 88.71% to 97.85 and 90.62% to 98.24% respectively. 2DCNN has shown the least performance on the dataset with real-world noise at different levels of SNR. CNN-GRU model has achieved better performance than other models on the dataset with real-world noise at different SNR level.

Table 3. Speaker identification performance of deep learning models using VoxCeleb1 dataset with real world noises

Model	Accuracy (%) at SNR					
	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
2DCNN	71.48	83.52	86.42	87.28	88.71	90.62
CNN-LSTM	86.39	89.63	93.63	96.11	96.84	97.14
CNN-BiLSTM	87.38	91.31	94.51	96.22	96.97	97.35
CNN-GRU (Proposed)	91.33	94.03	95.76	97.31	97.85	98.24

Figure 8, illustrates the comparison of speaker identification performance of different types of deep learning models on the datasets with the real-world noise at SNR ranges from -5 to 20 dB. The results show that the proposed CNN-GRU have higher accuracy than others in all the SNR level. 2DCNN performs lower than other models in all the SNR levels. There are only small variations between the accuracy of CNN-LSTM and CNN-BiLSTM in most of the SNR levels. CNN-GRU has shown high improvement than other models in the high noise (i.e., at SNR -5 dB and 0 dB).

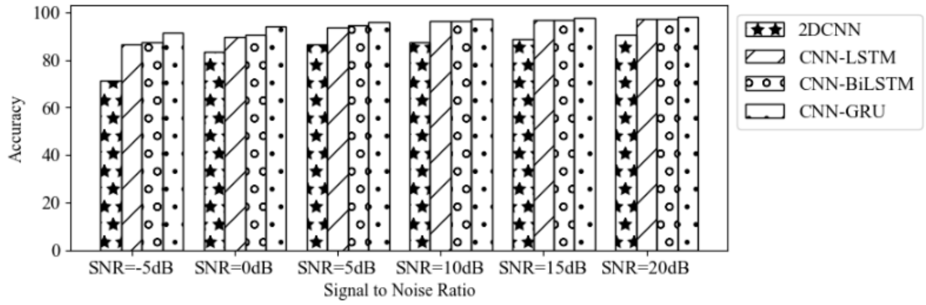


Figure 8. Speaker identification performance of models using dataset with real-world noises

Table 4, presents the speaker identification performance of the models on the VoxCeleb1 dataset without additive noise. The results show that the performance of the models ranges from 91.52% to 98.58%. 2DCNN has the least average accuracy which is 91.52% and CNN-GRU model has the highest average accuracy which is 98.58%.

Table 4. Speaker identification performance of the deep learning models using VoxCeleb1 without additive noise

Model	Accuracy (%)
2DCNN	91.52
CNN-LSTM	97.44
CNN-BiLSTM	97.79
CNN-GRU (Proposed)	98.58

Table 5, provides comparisons of the proposed model performance with the existing works to illustrate the efficiency of the proposed model. The existing works developed by using the deep learning model on the VoxCeleb1 dataset were selected for comparison. The result of the study proposed by Nagrani *et al.* [45], the study proposed by Kim and Park [46], and the study proposed by Ding *et al.* [47] were selected for comparison. The comparison shows that the proposed model has superior accuracy than the existing works. The speaker identification accuracy of the study on [45]–[47] are 92.10%, 95.30 and 96.01% respectively. The proposed model in this study achieved an average accuracy of 98.58%, which outperformed the existing works in speaker identification.

Table 5. Comparison of CNN-BiGRU based speaker identification performance with existing works

Methods	Dataset	Accuracy (%)
Nagrani *et al.* [45]	VoxCeleb1	92.10
Kim and Park [46]	VoxCeleb1	95.30
Ding *et al.* [47]	VoxCeleb1	96.01
CNN-GRU (Proposed)	VoxCeleb1	98.58

Generally, the proposed model performed better than other models experimented with in this study on the dataset with WGN and real-world noises at different levels of SNR. The proposed model has also performed better than other models in this study on the dataset without additive noises. The comparison results also confirmed that our model has better performance than existing works. The main reason for the improvement of the performance is that the proposed model has integrated the advantages of both 2DCNN and GRU architectures. Moreover, using cochleogram of the speech as an input improved the efficiency of the models under noisy conditions.

5. CONCLUSION

In this paper, a hybrid CNN and GRU model using cochleogram input is proposed for speaker identification under noisy conditions. Cochleograms were generated from each input utterance to feed into the model. The model integrated the advantages of both types of deep learning models (i.e., CNN and GRU models) and the cochleogram features. CNN models have advantages in automatically extracting short-term feature dependencies, few parameters for model training and adaptive learning from the parameters. The GRU models have advantages in extracting long-term correlation between features of the speaker, handling gradient vanishing and exploding problems, and converging faster during training. Cochleogram features have advantages in classifying speakers under noisy conditions because of the rich acoustic features of speakers and it has a finer resolution for speech with low frequency. In the proposed model, two CNN layers were employed first and two GRU layers of 256 cells were followed. The experiments were conducted on the VoxCeleb1 dataset with WGN, real-world noises (i.e., babble, restaurant and street) and without additive noise. The WGN and randomly selected real-world noises were added to the VoxCeleb1 dataset at the SNR of -5 dB to 20dB in the interval of 5 dB. To show the effectiveness of the proposed model, additional experiments were conducted by using the models 2DCNN, CNN-LSTM and CNN-BiLSTM. The model's performance was evaluated by using the accuracy. The accuracy of the proposed model was compared with the results of existing works. The experiment results show that the proposed model (i.e., CNN-GRU model) has better performance than other models. The comparison also indicates that the proposed model has better performance than existing works. The main reason for the performance improvements is that the proposed model integrated the advantages of the CNN model, GRU model and Cochleogram features. In the future, this study can be extended by using complex hybrid deep learning models. Further studies can also be conducted using a fusion of features as an input for small datasets.

ACKNOWLEDGEMENTS

This work was funded by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Grant No. GRANT4258). The authors acknowledge the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research at King Faisal University, Saudi Arabia, for financial support. The authors also acknowledge Adama Science and Technology University, Ethiopia, for financial support.

REFERENCES

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, Jan. 2010, doi: 10.1016/j.specom.2009.08.009.

[2] S. Ahmed, N. Mamun, and M. A. Hossain, "Cochleagram based speaker identification using noise adapted CNN," Nov. 2021, doi: 10.1109/ICEEICT53905.2021.9667916.

[3] A. G. Adami, "Modeling prosodic differences for speaker recognition," *Speech Communication*, vol. 49, no. 4, pp. 277–291, Apr. 2007, doi: 10.1016/j.specom.2007.02.005.

[4] K. Selvan, A. Joseph, and K. K. Anish Babu, "Speaker recognition system for security applications," in *2013 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2013*, Dec. 2013, pp. 26–30, doi: 10.1109/RAICS.2013.6745441.

[5] K. J. Han, M. K. Omar, J. Pelecanos, C. Pendus, S. Yaman, and W. Zhu, "Forensically inspired approaches to automatic speaker recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2011, pp. 5160–5163, doi: 10.1109/ICASSP.2011.5947519.

[6] F. Alegre, G. Soldi, N. Evans, B. Fauve, and J. Liu, "Evasion and obfuscation in speaker recognition surveillance and forensics," Mar. 2014, doi: 10.1109/IWBF.2014.6914244.

[7] N. Singh, R. A. Khan, and R. Shree, "Applications of speaker recognition," *Procedia Engineering*, vol. 38, pp. 3122–3126, 2012, doi: 10.1016/j.proeng.2012.06.363.

[8] L. Li *et al.*, "CN-Celeb: Multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77–91, Feb. 2022, doi: 10.1016/j.specom.2022.01.002.

[9] A. Kanervisto, V. Vestman, M. Sahidullah, V. Hautamaki, and T. Kinnunen, "Effects of gender information in text-independent and text-dependent speaker verification," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Mar. 2017, pp. 5360–5364, doi: 10.1109/ICASSP.2017.7953180.

[10] L. Chowdhury, H. Zunair, and N. Mohammed, "Robust deep speaker recognition: Learning latent representation with joint angular margin loss," *Applied Sciences (Switzerland)*, vol. 10, no. 21, pp. 1–17, Oct. 2020, doi: 10.3390/app10217522.

[11] S. Paulose, D. Mathew, and A. Thomas, "Performance evaluation of different modeling methods and classifiers with MFCC and IHC features for speaker recognition," *Procedia Computer Science*, vol. 115, pp. 55–62, 2017, doi: 10.1016/j.procs.2017.09.076.

[12] M. El Ayadi, A. K. S.O. Hassan, A. Abdel-Naby, and O. A. Elgendy, "Text-independent speaker identification using robust statistics estimation," *Speech Communication*, vol. 92, pp. 52–63, Sep. 2017, doi: 10.1016/j.specom.2017.05.005.

[13] M. India, P. Safari, and J. Hernando, "Self multi-head attention for speaker recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Sep. 2019, vol. 2019-September, pp. 4305–4309, doi: 10.21437/Interspeech.2019-2616.

[14] T. Kumar and R. K. Bhukya, "Mel spectrogram based automatic speaker verification using GMM-UBM," Dec. 2022, doi: 10.1109/UPCON56432.2022.9986424.

[15] J. C. Wang, C. Y. Wang, Y. H. Chin, Y. T. Liu, E. T. Chen, and P. C. Chang, "Spectral-temporal receptive fields and MFCC

balanced feature extraction for robust speaker recognition," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4055–4068, Feb. 2017, doi: 10.1007/s11042-016-3335-0.

[16] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors," *Speech Communication*, vol. 55, no. 2, pp. 237–251, Feb. 2013, doi: 10.1016/j.specom.2012.08.007.

[17] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," Dec. 2010, doi: 10.1109/ICSPCS.2010.5709752.

[18] Z. Weng, L. Li, and D. Guo, "Speaker recognition using weighted dynamic MFCC based on GMM," in *Proceedings - 2010 International Conference on Anti-Counterfeiting, Security and Identification, 2010 ASID*, Jul. 2010, pp. 285–288, doi: 10.1109/ICASID.2010.5551341.

[19] R. A. Malik, C. Setianingsih, and M. Nasrun, "Speaker recognition for device controlling using MFCC and GMM algorithm," Nov. 2020, doi: 10.1109/ICECIE50279.2020.9309603.

[20] R. Ajgou, S. Sbaa, S. Ghendir, A. Chamsa, and A. Taleb-Ahmed, "Robust remote speaker recognition system based on AR-MFCC features and efficient speech activity detection algorithm," in *2014 11th International Symposium on Wireless Communications Systems, ISWCS 2014 - Proceedings*, Aug. 2014, pp. 722–727, doi: 10.1109/ISWCS.2014.6933448.

[21] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, May 2013, pp. 7204–7208, doi: 10.1109/ICASSP.2013.6639061.

[22] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1791–1801, Aug. 2011, doi: 10.1109/TASL.2010.2101594.

[23] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Transactions on Multimedia*, vol. 14, no. 6, pp. 1684–1689, Dec. 2012, doi: 10.1109/TMM.2012.2199972.

[24] B. Ayoub, K. Jamal, and Z. Arsalane, "Gammatone frequency cepstral coefficients for speaker identification over VoIP networks," Mar. 2016, doi: 10.1109/IT4OD.2016.7479293.

[25] H. Wang and C. Zhang, "The application of gammatone frequency cepstral coefficients for forensic voice comparison under noisy conditions," *Australian Journal of Forensic Sciences*, vol. 52, no. 5, pp. 553–568, Mar. 2020, doi: 10.1080/00450618.2019.1584830.

[26] H. Choudhary, D. Sadhya, and V. Patel, "Automatic speaker verification using gammatone frequency cepstral coefficients," in *Proceedings of the 8th International Conference on Signal Processing and Integrated Networks, SPIN 2021*, Aug. 2021, pp. 424–428, doi: 10.1109/SPIN52536.2021.9566150.

[27] A. Torfi, J. Dawson, and N. M. Nasrabadi, "Text-independent speaker verification using 3D convolutional neural networks," in *Proceedings - IEEE International Conference on Multimedia and Expo*, Jul. 2018, vol. 2018-July, doi: 10.1109/ICME.2018.8486441.

[28] S. Shon, H. Tang, and J. Glass, "Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model," in *2018 IEEE Spoken Language Technology Workshop, SLT 2018 - Proceedings*, Dec. 2019, pp. 1007–1013, doi: 10.1109/SLT.2018.8639622.

[29] F. Ye and J. Yang, "A deep neural network model for speaker identification," *Applied Sciences (Switzerland)*, vol. 11, no. 8, Apr. 2021, doi: 10.3390/app11083603.

[30] S. E. Eskimez, P. Soufleris, Z. Duan, and W. Heinzelman, "Front-end speech enhancement for commercial speaker verification systems," *Speech Communication*, vol. 99, pp. 101–113, May 2018, doi: 10.1016/j.specom.2018.03.008.

[31] C. Liu, Y. Yin, Y. Sun, and O. K. Ersoy, "Multi-scale ResNet and BiGRU automatic sleep staging based on attention mechanism," *PLoS ONE*, vol. 17, no. 6 June, Jun. 2022, doi: 10.1371/journal.pone.0269500.

[32] S. Farsiani, H. Izadkhah, and S. Lotfi, "An optimum end-to-end text-independent speaker identification system using convolutional neural network," *Computers and Electrical Engineering*, vol. 100, May 2022, doi: 10.1016/j.compeleceng.2022.107882.

[33] A. Ashar, M. S. Bhatti, and U. Mushtaq, "Speaker identification using a hybrid CNN-MFCC approach," Mar. 2020, doi: 10.1109/ICETST49965.2020.9080730.

[34] A. M. Jalil, F. S. Hasan, and H. A. Alabbasi, "Speaker identification using convolutional neural network for clean and noisy speech samples," in *1st International Scientific Conference of Computer and Applied Sciences, CAS 2019*, Dec. 2019, pp. 57–62, doi: 10.1109/CAS47993.2019.9075461.

[35] S. Soleymani, A. Dabouei, S. M. Iranmanesh, H. Kazemi, J. Dawson, and N. M. Nasrabadi, "Prosodic-enhanced siamese convolutional neural networks for cross-device text-independent speaker verification," Oct. 2018, doi: 10.1109/BTAS.2018.8698585.

[36] D. Salvati, C. Drioli, and G. L. Foresti, "A late fusion deep neural network for robust speaker identification using raw waveforms and gammatone cepstral coefficients," *Expert Systems with Applications*, vol. 222, Jul. 2023, doi: 10.1016/j.eswa.2023.119750.

[37] G. Costantini, V. Cesarini, and E. Brenna, "High-level CNN and machine learning methods for speaker recognition," *Sensors*, vol. 23, no. 7, Mar. 2023, doi: 10.3390/s23073461.

[38] 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.

[39] W. Lambamo, R. Srinivasagan, and W. Jifara, "Analyzing noise robustness of cochleogram and mel spectrogram features in deep learning based speaker recognition," *Applied Sciences (Switzerland)*, vol. 13, no. 1, Dec. 2023, doi: 10.3390/app13010569.

[40] Z. Zhao *et al.*, "A lighten CNN-LSTM model for speaker verification on embedded devices," *Future Generation Computer Systems*, vol. 100, pp. 751–758, Nov. 2019, doi: 10.1016/j.future.2019.05.057.

[41] M. Bader, I. Shahin, A. Ahmed, and N. Werghi, "Hybrid CNN-LSTM speaker identification framework for evaluating the impact of face masks," in *2022 International Conference on Electrical and Computing Technologies and Applications, ICECTA 2022*, Nov. 2022, pp. 118–121, doi: 10.1109/ICECTA57148.2022.9990138.

[42] H. S. Das and P. Roy, "A CNN-BiLSTM based hybrid model for Indian language identification," *Applied Acoustics*, vol. 182, Nov. 2021, doi: 10.1016/j.apacoust.2021.108274.

[43] Y. H. Liu, X. Liu, W. Fan, B. Zhong, and J. X. Du, "Efficient audio-visual speaker recognition via deep heterogeneous feature fusion," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10568 LNCS, Springer International Publishing, 2017, pp. 575–583.

[44] B. C. Mateus, M. Mendes, J. T. Farinha, R. Assis, and A. M. Cardoso, "Comparing LSTM and GRU models to predict the condition of a pulp paper press," *Energies*, vol. 14, no. 21, Oct. 2021, doi: 10.3390/en14216958.

[45] A. Nagraniy, J. S. Chungy, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Aug. 2017, vol. 2017-August, pp.

2616–2620, doi: 10.21437/Interspeech.2017-950.

[46] S. H. Kim and Y. H. Park, “Adaptive convolutional neural network for text-independent speaker recognition,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Aug. 2021, vol. 1, pp. 641–645, doi: 10.21437/Interspeech.2021-65.

[47] S. Ding, T. Chen, X. Gong, W. Zha, and Z. Wang, “AutoSpeech: Neural architecture search for speaker recognition,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Oct. 2020, vol. 2020-October, pp. 916–920, doi: 10.21437/Interspeech.2020-1258.

BIOGRAPHIES OF AUTHORS

Wondimu Lambamo received the BSc. Degree in computer science and information technology from the Jigjiga University, jigjiga, Ethiopia, in 2010. He received the MSc. degree in computer science from Arba Minch University, Arba Minch, Ethiopia, in 2017. He is currently pursuing Ph.D. degree with Adama Science and Technology University, Adama, Ethiopia. His current research interest includes natural language processing, signal, image processing, biometrics and deep learning. In the past, he was lecturer of computer science with Wollega University. He is currently a lecturer, researcher and department head of computer science with Wachemo University. He published a number of research articles with the recognized journals. wondimuwcu@gmail.com.

Ramasamy Srinivasagan Graduated from the Madurai Kamaraj University, India in Electrical and Electronics Engineering. He holds M.E degree in Electronics and Communication – specialised in VLSI Design and a Ph.D. in Mixed Signal VLSI systems from National Institute of Technolgy, Trichy, Tamilnadu, India. He is Faculty in Computer Engineering, King Faisal University. His research interests include Deep Learning, Tiny Machine Learning, VLSI Systems, Industry 5.0 and systems-on-chip. rsamy@kfu.edu.sa.

Worku Jifara received the Bacheleor degree in Mathematics from Madawalabu Universit, Robe, Ethiopia, Masters degree in AppliedMathematics from Addis Ababa Univeristy, Addis Ababa, Ethiopia, and Ph.D. degree in Computer Science and Technology from Harbin Institute of Technology, Harbin, China. Currently, he is an assistant professor and researcher of computer science with Adama Science and Technology University. He is also an ICT director of Adama Science and Technology University. His research intrest includes medical image processing, big data analysis, pattern recognition and satellite image processing. worku.jifara@gmail.com.

Ali Alzahrani received the B.E degree in Computer Engineering from Umm Al-Qura University, Mecca, Saud Arabia, and the M.Sc. and Ph.D. degrees in computer engineering from the University of Victoria, BC, Canada, in 2015 and 2018, respectively. He is currently an Associate Professor with the Department of Computer Engineering, King Faisal University. His research interests include hardware security, speech processing, Deep Learning encryption processors, image processing, and systems-on-chip. aalzahrani@kfu.edu.sa.