Multimedia Tools and Applications (2023) 82:33111–33133
https://doi.org/10.1007/s11042-023-14942-9

A deep learning approach for text-independent speaker recognition with short utterances

Rania Chakroun [1,2] · Mondher Frikha [1,3]

Received: 11 January 2022 / Revised: 30 June 2022 / Accepted: 22 February 2023 /
Published online: 6 March 2023

Abstract
Recently, the speaker recognition techniques have been widely attractive for their extensive use in many fields, such as speech communications, domestic services, security and access control and smart terminals. Today's interactive devices like smart-phone assistants and smart speakers need to deal with short duration speech segments. However, existing speaker recognition applications perform poorly when short utterances are available and require relatively long speech to perform well. Aiming at solving this problem, we introduce in this paper, a novel method to enhance the speaker recognition capability with short utterance speaker recognition applications. For this purpose, we considered new deep neural network architectures based on convolutional neural network (CNN) and recurrent neural network (RNN). The proposed method is evaluated with the standard i-vector based on Probabilistic Linear discriminant analysis (PLDA) approach. The experimental results show that our model could outperform the i-vector -PLDA baseline system and enhance the speaker recognition capability when significant and short utterance duration are used.

Keywords Speaker recognition · i-vector · PLDA · Short utterances · Deep learning · DNN · CNN · LSTM · RNN

1 Introduction

Biometric identification or biometrics refers to the process of recognizing an individual based on his distinguishing characteristics. It consists on methods for uniquely recognizing humans

✉ Rania Chakroun
chakrounrania@yahoo.fr

Mondher Frikha
mondher@yahoo.fr

[1] Advanced Technologies for Image and Signal Processing (ATISP) Research Unit, Sfax, Tunisia

[2] National School of Engineering of sfax, Sfax, Tunisia

[3] National School of Electronics and Telecommunications of Sfax, Sfax, Tunisia

based on one or more intrinsic physical or behavioural traits [34]. In fact, Biometric identification provides high level of security compared to Traditional methods involving keys, passports, smartcards, user ID, PIN numbers and passwords which can be easily stolen or forged. Nowadays, there are many biometric technologies based on the physiological characteristics such as face, fingerprint, iris and behavioural characteristics like hand written signature, gait and keystroke [18]. Hence, a biometric system operates by acquiring biometric data from an individual, extracting the appropriate features and comparing it with the models set in the database.

Nowadays, the voice is considered as an important human trait most take for granted in natural human-to-human interaction and communication [37]. Speaking to someone over a telephone usually begins by the identification of who is speaking and in cases of familiar speakers, a subjective verification is needed by the listener to ensure that the identity is correct so that the conversation can proceed.

Automatic Speaker recognition is the process of recognizing the appropriate individual only from his voice. This technique make it possible the use of the speaker's information includes in the speech waves to verify the user identity of and control the access to many services including voice dialling, banking by telephone, database access services, telephone shopping, voice mail, information services, remote access to the computers, transaction security of bank trading and remote payment,…etc.

Speaker recognition can be divided into two main applications which are speaker identification and speaker verification. In speaker identification, human speech of an individual is used to identify who that individual is among a set of speaker models. In fact, the speech from an unknown speech utterance is compared against each of the trained speaker models and the best matches is the identified speaker. In speaker verification, human speech of an individual is used to verify the claimed speaker identity of that individual.

Both speaker identification and speaker verification tasks can be text-dependent or text independent. In text-dependent systems, the speaker must say a specific phrase or words for both training and testing phases. This method is simpler to the system however it cannot be efficient since it is limited by a specific predetermined speech. While in text-independent systems, the system identifies the speaker from any spoken phrase regardless of the utterance content. Text-independent speaker recognition is more complex to handle for the system but it is more flexible for the users since there are no limitations for the text used in the test or in the train phases and the speaker must be recognized independent of what is saying [9].

Nowadays, Short Utterance Speaker Recognition is becoming a major consideration of modern speaker recognition research. State of the arts speaker recognition methods need a large amount of speech data for training speaker models. However, in real wold circumstances, it is difficult to acquire a large amount of appropriate speech data. In fact, most of the times background noise gets into the way. Also, a faulty recording reduces most of the speech voice, leaving behind only few seconds of intelligible speech. Nevertheless, the interest in speech and speaker recognition applications over fixed telephone, mobile phone and hand-held palm devices has been augmented. These devices are almost used in adverse environments such as city streets, airports, offices and cars,…, etc. The amount of required speech is therefore affected. For that, it becomes necessary to take into account the speaker specific information from short utterances of speech, so that speaker recognition systems should be performed even when there is only a few amount of speech data available.

The development of a realistic speaker recognition system can't be complete without taking into account of the problems related to the memory and computational resource limitation.

Hence, the system should be performed with the minimum as possible of speech utterance durations. In this context, the use of short utterance speaker identification is essential to develop an efficient application.

In order to deal with high performing application, state of the art Speaker recognition systems rely on significant amount of speech for enrolment and testing [43]. In fact, traditional methods like GMMs, SVMs and even conventional i-vectors need sufficient amount of speech data (>15 s) for the extraction of sufficient statistics to build the speaker models [46]. The performances of such systems decrease in short-utterance conditions. Even though, in real-world scenarios, it is recommended that accurate speaker recognition application be performed using short segments of speech duration. Indeed, with the emergence of voice based interactive devices like the smart phone assistants, smart home devices like smart speakers, vehicles, banking applications, distant navigation and control systems, it is imperative to use short duration speech segments [27]. So that, speaker recognition systems integrated into such devices will be more suitable in terms of time and memory complexity of the systems, real world circumstances imposed by the speaker itself, the environmental conditions and even the quality of the transmission channel and background noise that limit the acquisition of sufficient quantity of speech data. That is why current trends for speaker recognition are addressed for searching suitable and efficient systems to tackle the problem of speaker recognition using short duration segments [2, 74]. Despite that, existing speaker recognition solutions are still limited and the efforts made need to be more improved.

Based on our survey in speaker recognition domain, with particular emphasis on short utterance speaker recognition, we notice that there is a lack of a particular focus on the problem related to speaker identification based on short utterances [20, 45, 51]. In this study, we pay a specific attention to speaker identification task when a little amount of speech is available. We focus on such system taking into account that there is no restriction concerning the text content of the input speech data, only a little amount of speech is available for both training and testing phases.

In this context, such research problems require to deal with Artificial Intelligence (AI). In fact, AI is the study of complex information processing problems that often have their roots in some aspect of biological information processing. The goal of the subject is to identify interesting and solvable information processing problems, and solve them [59]. In other words, AI is a general term that require the use of a computer to model an intelligent behaviour, nearly human-level understanding of the data, with minimal human intervention. In this field, Deep Learning, allows computational models which are composed of multiple processing layers in order to learn representations of data with multiple levels of abstraction. These methods have dramatically improved many domains. Therefore, Noticing the increasing use of Deep Neural Networks (DNNs) models that performed recently well in pattern recognition tasks such as speech recognition [80], face recognition [26] and also in speaker recognition domain [5, 78] which have shown outperformance compared to i-vector based systems [73]. In fact, the i-vector based on probabilistic linear discriminant analysis (PLDA) speaker recognition systems are considered as recent state of the art systems [8, 40, 44]. Most recent works depend on using the i-vector based on PLDA technique as a baseline system for this field [4, 8, 40, 44, 48, 49, 54, 62, 85].

We propose in this study, a novel DNN architecture for speaker identification task using short utterance duration. Our proposed model is applied with both Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) models that have shown good performance in some speaker recognition works [5], and we want to test their performance

in short utterance conditions. We propose then a Deep Neural Network (DNN) architecture based on the CNN model [55] and another system based on the RNN [1] with the multimodal Long Short-Term Memory (LSTM) model [1] for the speaker identification task. These systems are carried out using the proposed Cepstral Mean and Variance Normalized coefficients (CMVNC) that use the Cepstral Mean and Variance Normalisation (CMVN) technique [81] that we explain in the following section. We further compare the performances of both proposed systems against speaker recognition systems using CNN models and speaker recognition system using LSTM RNN based on the standard Mel frequency cepestral cofficients (MFCC) [81]. The compared results among the different models evaluated with speakers taken from two different datasets, NIST SRE 2010 dataset [64] and VoxCeleb2 dataset [11], are described and explain the outperformance of the proposed systems when a little amount of speech (<15 s) is used for data evaluation. We systematically analyse the effect of duration of speech utterances on speaker identification with the proposed architecture and we try to overcome the problem of short utterances through the use of the new proposed CMVNC features that import supplementary information detected from the speech signal and improve the system performance.

The rest of this paper is organized as follows. In the next section, a short survey about previous works in short utterance speaker recognition and the use of the DNN approach in speaker recognition domain is given. In Section 3, the proposed speaker identification approach is described, experimental set-up and results are demonstrated and discussed in Section 4 and conclusions are drawn in Section 5.

2 Related works

Nowadays, researchers are starting to apply speaker recognition in many applications. With the emergence of voice based interactive devices like the smart phone assistants, the increase use of Internet [32], distant navigation and control systems, it is imperative to verify the identity of users to accomplish the appropriate tasks. We can also cite the importance of integrating the task of speaker recognition in indexing multimedia data and facilitating the research of appropriate documents especially with the explosion of the amount of information [82] and multimedia data [29] over past decades. However, it is highly difficult to collect a sufficient amount of speech data in most situations. For example, in forensic applications or even with the presence of environmental circumstances, the speech obtained could be broken, unclear, or recorded in noisy conditions or contains some breaks and a little amount of real speech. Furthermore, most users are reluctant to provide much speech data especially in the test phase like in telephone banking application. Some other conditions could be imposed by the state of health of the speaker, his character,…, etc. Besides, realistic applications can impose several constraints related to the system itself. For example, the problem related to the memory and computational resource limitation or even the utterance duration fixed by the system. These entire conditions prevent the collect of a large amount of data as required by conventional speaker recognition approaches and make short utterance speaker recognition arises as an important area of research in such cases.

Over the last decades, a lot of methods have been proposed and investigated for speaker recognition purposes [60, 70, 77]. In fact, a great progress has been made in the task of automatically recognizing the identities through the voice of persons and most successful state of the art applications are based on the well-known Gaussian Mixture Model (GMM) [47, 71]

and the GMM-Universal background model(GMM-UBM) [66, 72]. The results indicate that these models provide a robust speaker representation and they achieve for example in [71] high recognition accuracy that exceeds 90% with clean speech and more than 80% with more than 30s of training speech utterance duration using telephone speech with a 49 speaker population.

The use of the SVM [17, 28, 63] and supervectors [45] is also promoting in accurately describe the speakers. More recently, Joint Factor Analysis (JFA) [14] and i-vector models [15] have also been investigated.

The research methodologies in the area of speaker recognition have provided high recognition performance with sufficient amount of speech data [53, 55]. The efficacy of most of these the state-of-the-art methodologies degrades considerably when a little amount of speech (<15 s) is used for data evaluation [36, 46]. In fact, the authors reported in [83] that when the test speech was shortened from 20 seconds to 2 seconds, the performance degraded sharply in terms of equal error rate (EER) from 6.34% to 23.89% on a NIST SRE database. Also, the authors show in [58] that when the length of the test speech is less than 2 seconds, the EER raised to as high as 35.00%.

That's why short utterance speaker recognition remains a focus of interest of many researches for quite some time [51]. For example, with the GMM speaker recognition technology, which uses the segmented statistical features of the speech spectrum to recognize the speaker, it is difficult to obtain good results in short utterance conditions. Even so, the performance of the state-of-the-art technologies degrades drastically and demonstrate considerable limitations with the short-term speech in the spectral statistics [13, 67]. Also in [51], the recognition performance seems good and achieve 96% with 10s of test speech utterances and fall down to 79% with 3 s of test speech duration. However, the training phase is dealt with 100 s of speech data duration which is a considerable amount of speech data and only test speech utterances are considered as short with only 11 male talkers.

The short utterance was recently considered as an open challenge to the research community so that numerous attempts have been made to mitigate this issue. The relevant works concentrated on different aspects of Automatic Speaker recognition like feature extraction techniques [3, 52], speaker modelling techniques [39, 42], phonetic information [53, 79], score normalisation techniques [31], etc. to compensate the limited duration issue. Even though, further attention is paid on i-vector and probabilistic linear discriminant analysis (PLDA) based speaker recognition systems which were considered as recent state of the art systems due to their importance in improving the existing systems performance with the reduction of utterance lengths [40, 44]. In fact, the use of the proposed system in [44] shows over 10% improvement in EER over the baseline system.

With the rise of deep learning technology, which realized large performance improvements in many other pattern recognition tasks such as speech recognition [12] and face recognition [76], the use of DNNs is involved in speaker recognition domain and succeed to achieve comparable results with existent successful methods. In fact, DNNs were used for speaker recognition to replace or improve state-of-the-art i-vector based on PLDA system. For example, works in [50], has shown promising results by using DNN acoustic models instead of Gaussian mixture models to extract sufficient statistics [50]. Improvement from the proposed framework compared to a state-of-the-art system are of 30% relative at the equal error rate when evaluated on the telephone conditions from the 2012 NIST SRE dataset. Also, the use of DNN bottleneck features instead of conventional MFCC features [56] lead to better performance with the system. The use of DNNs for complementing PLDA in [6, 65] or to

replace it [23]. For example, the authors show in [6] that the inclusion of deep neural network performs better than the PLDA baseline, achieving an equal error rate of 2.92% as compared to 3.37%.

Another attempts are made with end-to-end systems that have been proven to be competitive for text-independent speaker Recognition with short test utterances and an abundance of training data [73, 74]. Other recent research focus on using DNN based speaker recognition systems with more difficult conditions such as distant talking [87]. Later on, recurrent neural networks (RNN) which have been utilized in a number of studies were employed for speaker recognition task [35, 84] with a long short-term memory (LSTM) architecture and using Mel-Frequency Cepstrum Coefficients (MFCC) as inputs. In fact, in [84], the proposed system succeed to improve the recognition performance and achieve more than 70% of speaker recognition accuracy on VoxCeleb2 dataset using LSTM recurrent neural networks implemented with 19-dimensional MFCCs, their first and second derivatives, along with the first and second derivatives of energy.

Many other studies have utilized the CNN-based models for speaker recognition [62, 86]. For example, in [62] a CNN architecture was introduced, which outperformed the I-vector-based methods. For identification, the proposed system achieve an 80.5% classification accuracy which is higher than traditional state of the art baseline system with VoxCeleb database. Accordingly, deep learning of the voiceprint using deep neural networks may solve the speaker recognition problem with an adequate system performance that can also deal with short utterance conditions. In fact, the attempts of using DNN in SUSR conditions occur recently [41, 43, 53, 74]. Even so, works are still limited and the achieved improvements remained related to the speech duration employed for both training and testing tasks, the questioned task of speaker recognition (verification, identification) [70], the number of speakers used, and so on. Thus, extra works are required to search for of effective methods for the speaker identification task when a limited amount of speech is only available.

In order to address this research problem, which is now becoming a major consideration of modern speaker recognition applications [2, 74], we investigate in this paper the DNN approach to improve the performance of speaker identification system under short utterance evaluation conditions. Since short utterances are likely to contain less speaker characteristics and information compared to long utterances, we hypothesize that it is beneficial to evaluate and take advantage from the convolutional neural network (CNN) models and also the LSTM network together with the RNN models to obtain more adapted classifier that can well capture all the variations present in the short speech utterances. Thus, we will adopt two DNN systems architecture and we will reinforce them with new well adapted features in order to compensate the limit caused by the reduction of data duration.

3 Methodology

This article describes the development of a speaker identification system based on Deep learning algorithms that provide acceptable performance under constrained operating conditions dealing with limited data duration. Initially, an acoustic pretreatment is done and feature extraction is performed using MFCC coefficients to obtain the required features. Furthermore, we propose to use normalized cepstral coefficients which are the CMVNC features. We give the adequate description of these features in the following subsection. These coefficients are more adapted to the signal and time variations since they normalize distribution parameters of

cepstral coefficients over specified time interval using sliding windows which increase the robustness of the system against the effects of linear channel and slowly varying additive noise. Hence, we improve the ability of the system to capture the maximum information with limited speech data duration. Thereafter, the proposed CNN models and the LSTM-RNN classification models are used to perform speaker recognition. The flow diagram of the first proposed system using the CNN models is shown in Fig. 2. The second proposed system based on the LSTM-RNN models is given with Fig. 3. For ensuring the effectiveness of the present work, the speaker accuracy is calculated and compared against baseline system using the standard i-vector based on PLDA technique that we present with Fig. 1.

3.1 Acoustic features

The acoustic pretreatment phase includes feature extraction process that represent one of the most important steps in the speaker recognition systems as it extracts the best parametric representation of the acoustic signals. The most commonly used features in speaker recognition are MFCC as the Mel frequency bands approximates human hearing perceptions of sounds more closely than any other systems. In fact, the state-of-the-art systems use many kinds of features were the most successful and popular are Mel-Frequency Cepstral Coefficients (MFCCs) [81]. Many recent works assert the adoption of the MFCC features as input with Deep Neural Networks architecture for speech and speaker recognition systems [19, 61]. In fact, the most important step for any speech recognition system is to extract the features that are good at finding linguistic content and discards all other unwanted information like noise, emotions, etc. [19]. That's why choosing the input data is essential for speaker recognition domain. In this study, these features are extracted from a 25 ms hamming window with 10 ms overlap. 12 MFCC coefficients together with log-energy were calculated every 10 ms and augmented with their first and second derivatives leading to 39-dimensional feature vector per frame. In fact, this structure of feature vector is widely used in the state-of-the-art speaker recognition systems and also in recent works [7, 81, 88].

We tried to reduce the effect of the variability of the extracted characteristics from the speech signal from a session to another. Thus, we used the Cepstral Mean Normalization (CMN) [81]. We adopt another kind of normalization which is the Cepstral Mean and Variance Normalization (CMVN) [81] to improve robustness. We use then the CMVNC

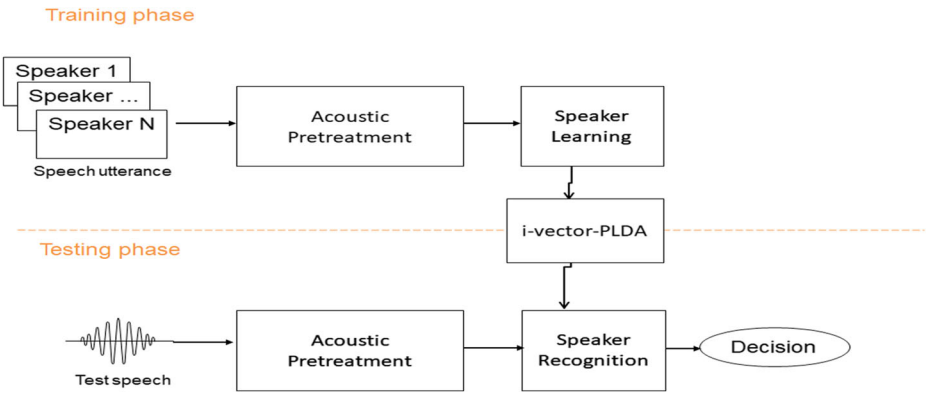

Fig. 1 The baseline Short Utterance Speaker Recognition system using the i-vector based on PLDA technique

coefficients: Mean and Variance Normalized MFCC, which is a short-time cepstral representation of a speech in which we normalize the feature vector coefficients using the CMVN technique. In fact, for a given feature vector $X = \{x[1], x[2], \ldots, x[N]\}$ of MFCC coefficients, the resulting feature vector presenting the MVNMFCC coefficients is calculated as follows:

$$\widehat{x}[n] = \frac{x[n]-\overline{x}}{\sigma_x} \tag{1}$$

$$\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x[n] \tag{2}$$

$$\sigma_x^2 = \frac{1}{N}\sum_{n=1}^{N} (x[n]-\overline{x})^2 \tag{3}$$

Where N represents the number of MFCC coefficients in a feature vector and n is the order of the MFCC coefficient in this vector.

3.2 Modelling techniques

3.2.1 Convolutional neural network (CNN)

Recently, Deep Neural Network (DNN) architectures have been applied in many fields including computer vision, speech recognition, natural language processing, machine translation and bio-informatics [5]. A DNN is an artificial neural network dealing with extra hidden layers between the input and output layers, which permit to model complex data with fewer and expressive features.

Recently, the use of CNNs have received an increased attention from the research community and it have gradually become the main research tool in the field of image and speech [26, 80]. A CNN is a DNN that imitates how the visual cortex of the brain processes and recognizes images. In fact, the main aim of CNN is to discover local structure in the input data. In speaker recognition domain, the spectrogram [57] gives a large amount of information about the personal characteristics of the speaker. It permits to dynamically show the characteristics of the signal spectrum change. Thus, the spectrogram is considered as an effective tool for researchers to apply CNN with signal-based applications and the feature vectors need to be obtained through it. Although speech is a time-varying signal with complex correlations at a range of different timescales, the spectrogram provides a good solution to well visualize the different variations of the speech signal. In fact, the spectrogram is used as the input of the CNNs. It is a two-dimensional signal and it contains the identity information of the speaker. Thus, CNNs can provide translation invariance in time and space, so we can obtain the voiceprint features in the spectrogram space without destroying the time sequence. Therefore, speaker recognition study proposes to use the spectrogram as the input of the convolutional neural network.

The CNN model consists of convolutional layers followed by activation functions, pooling layers, fully connected layers and finally a classification stage. The inclusion of a dropout layer allows regularization for reducing overfitting [68].

Typically, in order to adjust the speech signals to be suitable for CNN classification, they are processed to obtain their spectrograms which is made by applying the Short Time Fourier Transform (STFT) to the speech segments. In this work, in order to take advantages of adequate acoustic features, the input spectrograms use MFCC and MVNMFCC features to

better represent the speech information belonging to each person in the speech signal dataset. In fact, the speech signals are divided into frames. Hamming windows is applied to each frame with window size of 25 ms with 10 ms overlap. Then, for each frame 39 cepstral features are extracted. In parallel, the spectrogram of each frame is generated. A Spectrogram is simply a signal strength versus time at different frequencies and is generated by applying STFT. At the end of this step, each speech signal is represented by spectrograms.

The structure of the proposed CNN model (Fig. 2) includes three convolutional layers with 3 maximum pooling layers. For the input layer, frames of 32-dimensional filter-bank features, which belong to the same person are grouped together as a feature map. Kernel size of each convolutional layer is 3 × 3, and the stride is set to be 2 × 2. The Relu activation function [25] is employed for activation process in the training phase. This function is an efficient activation function to be used in Deep Learning. Each convolutional layer is connected to pooling layer of 3 × 3 max pooling. Then, Batch Normalization [33] is used to get the speaker representation due to its robustness to internal covariate shifts. Then, flatten and dense layers are considered. The dense layer, with size of 64, is used for the classification task since we have 64 classes, and softmax function [21] is adopted as the scoring method.

3.2.2 Recurrent neural network (RNN)

In recent studies, the RNNs have achieved excellent results in language modelling tasks as outlined in [22, 75]. In fact, the RNN model has been a highly preferred method, especially for sequential data [10, 69]. Recent advances in deep learning have given rise to the use of sequence-to-sequence models for speaker recognition [16]. In simple terms, unlike traditional neural networks, RNN could use its reasoning about previous events to inform later ones and decide on the current input. Indeed, the output of the current time stamp depends on the previous time stamp. There is memory assigned to every cell of RNN. This memory keeps the track of previously computed outputs. In other terms, in RNN the decision made at time t − 1 affects the decision at time t. Thus, the decision of how the network will respond to new data is dependent on two things, the current input, and the output from the recent past. For a given input time series $x = \{x_1, x_2,..., x_T\}$, the RNN computes the hid-den state sequence denoted by $h = \{h_1, h_2,..., h_T\}$ and the output sequence y = $\{y_1, y_2,..., y_T\}$ by iteratively calculating the following two equations:

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{4}$$

$$y_t = g(W_{hy}h_t + b_y) \tag{5}$$

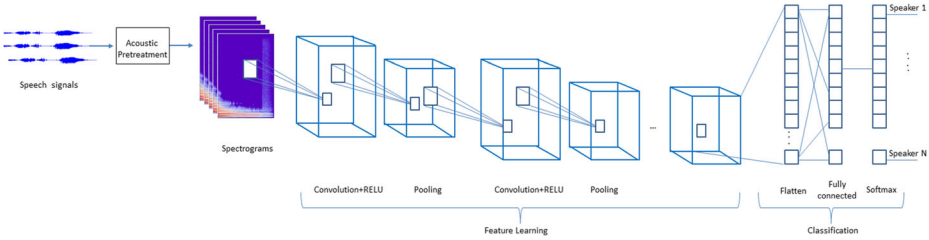

Fig. 2 The proposed Short Utterance Speaker Recognition system based on the CNN models

Where W represents the weight matrices, the vectors b_h and b_y denote the bias of the hidden layer and the output layer. The activation functions for the hidden layer and the output layer are represented by $f(.)$ and $g(.)$, respectively.

With RNN, the hidden state h_t at the time step t is used to memorize the network and captures all the information contained in the previous time steps. But, RNN network have vanishing and exploding gradient problem during back propagation which affects their performance. Therefore, they cannot carry out long-term dependencies of sequential data from earlier time steps to later ones. It suffers from short-term memory so that RNN network can forget what it has seen in longer sequences.

3.2.3 Long short-term memory (LSTM) RNN

As the common drawback of a traditional RNN model is the failure to store information for long period, the solution to the shortcomings of this model was by introducing LSTM networks [30]. LSTMs are a special Recurrent Neural Network with memory cells that allows the neural network to take long term dependency into account. In fact, a special memory cell architecture in LSTM makes it easier to capture information for long period. Thus, LSTM architecture can mitigate the vanishing problem and so it is suitable for the problem of long-term dependencies. In fact, for an input time series represented by $x = \{x_1, x_2,..., x_T\}$, the LSTM maps input time series to two output time sequences $h = \{h_1, h_2,..., h_T\}$ and y = $\{y_1, y_2,..., y_T\}$ iteratively by updating the states of memory cells. The steps procedures are given by following equations:

$$f_t = \sigma\left(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}C_{t-1} + b_f\right) \tag{6}$$

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}C_{t-1} + b_i) \tag{7}$$

$$U_t = tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \tag{8}$$

$$C_t = U_t i_t + C_{t-1}f_t \tag{9}$$

$$O_t = tanh(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}C_{t-1} + b_o) \tag{10}$$

$$h_t = O_t * tanh(C_t) \tag{11}$$

$$y_t = k\left(W_{yh}h_t + b_y\right) \tag{12}$$

Where input weight matrices are denoted by W_{ix}, W_{fx}, W_{ox} and W_{cx}, respectively. Recurrent matrices are represented with W_{ih}, W_{fh}, W_{oh}, and W_{ch}, respectively, and W_{yh}represents the hidden output weight matrix, and W_{ic}, W_{fc}, and W_{oc} represent the weight matrices of peephole connections. The vectors b_i, b_f, b_o, b_c, b_y, are the corresponding bias vectors. σ is the logistic sigmoid function, and i, f, o, and c are respectively the input gate, forget gate, output gate, and cell activation vectors. Tanh is the network output activation function, the ReLU function in our experiments.

In this work, we adopt a speaker recognition system based on the use of the LSTM RNN networks. The structure of our proposed system is given with the following Fig. 3.

4 Experimental results

We conduct our speaker recognition experiments using recently used corpus in this field [5, 38]. We use conversational telephone and microphone (phone call and interview) speech utterances extracted from the NIST SRE 2010 database [64]. This corpus contains English speech recordings from a large number of male and female speakers with multiple sessions per speaker. The core evaluation condition includes speech samples of varying lengths from telephone conversations, conversations recorded over a room microphone channel, and conversational speech from an interview scenario recorded over a room microphone channel. Some of the telephone conversations have been collected in a manner to produce particularly high, or particularly low, speaker vocal efforts. A detailed description of the data, tasks and rules of SRE10 can be found in the evaluation plan available in [64]. In this work, we consider 64 speakers from NIST SRE 2010 corpora that involve trials taken from both male and female speakers.

The second set of experiments are evaluated on voxCeleb2 database [11]. This dataset consists of speech segments from unconstrained open-source media like YouTube videos for several thousand individuals. The dataset is fairly gender balanced. The speakers span a wide range of different ethnicities, accents, professions and ages. Videos included in the dataset are shot in a large number of challenging visual and auditory environments. These include interviews from red carpets, out-door stadiums and quiet indoor studios, speeches given to large audiences, excerpts from professionally shot multimedia, and even crude videos shot on hand-held devices. The VoxCeleb2 corpus were acquired using an automatic pipeline based on computer vision techniques. For a full description of the pipeline and an overview of the datasets, see [11]. All noise, reverberation, compression and other artifacts in the dataset are natural characteristics of the original audio and have not been removed. Hence, the recognition of speakers across such varied conditions is representative of many challenges that are required to be taken into account today with real world applications. In order to well evaluate the systems performance with this corpus, the evaluations are also made with 64 male and female speakers from voxCeleb2 database.

In order to make a comparison with previous works [8, 81] and examine the performance of the proposed speaker identification system, we carry out experimental evaluations as follows:

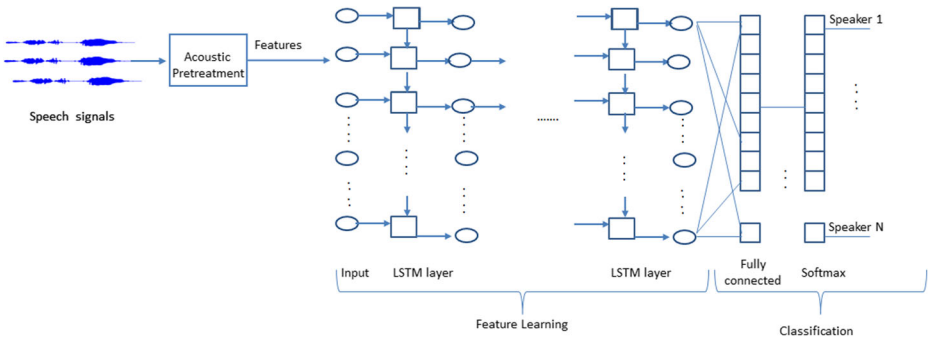

Fig. 3 The proposed Short Utterance Speaker Recognition system based on the LSTM-RNN models

we consider a baseline speaker identification system based on the i-vector-PLDA approach. Then, we consider a first proposed system based on the use of CNN models and another proposed system based on the LSTM-RNN approach. The different systems are implemented and evaluated under different scenarios. In fact, we use different training and testing speech utterances durations. We perform evaluations with utterances having a length of 15 s and 5 s for each speaker for the train task and speech segments having a length of 2 s 1 s and 0.5 s per speaker for the test task. We perform a first set of experiments with MFCC features. Then, we adopt CMVNC acoustic features for a second set of experiments to examine performances. The achieved results obtained from the different systems are compared so we can highlight the contribution of the proposed approach for short utterances speaker identification.

4.1 Speaker recognition with 15 s of training data duration

To evaluate the performance of the proposed system, we perform the first set of experiments using the baseline speaker identification system based on the i-vector-PLDA approach. Then, we consider a first proposed system based on the use of CNN models and another proposed system based on the LSTM-RNN approach. We use the MFCC coefficients as acoustic features extracted for the different systems. For the CNN model, the experiments are evaluated with a model consisting of the following layers. In fact, we use three convolutional layer, with kernel size of (3,3) and Rectified Linear Unit (ReLU) activation function. This function is an efficient activation function to be used in Deep Learning. Each convolutional layer is connected to a MaxPooling layer with pool size of (3,3) and followed by batch normalization. Then we used a Flatten layer and the outputs are fed into a dense layer having the same ReLU activation function. To avoid overfitting, we use a Dropout with $\alpha = 30$ so that the model will drop out 30% of weights. We finally used a fully connected layer with 64 neurons since the dataset is composed of 64 speakers, with an activation function being softmax.

For the CNN model, the experiments are evaluated with a model consisting of the following layers. In fact, we use three convolutional layer, with kernel size of (3,3) and Rectified Linear Unit (ReLU) activation function. This function is an efficient activation function to be used in Deep Learning. Each convolutional layer is connected to a MaxPooling layer with pool size of (3,3) and followed by batch normalization. Then we used a Flatten layer and the outputs are fed into a dense layer having the same ReLU activation function. To avoid overfitting, we use a Dropout with $\alpha = 30$ so that the model will drop out 30% of weights. We finally used a fully connected layer with 64 neurons since the dataset is composed of 64 speakers, with an activation function being softmax.

The LSTM-RNN consists of an input layer, two hidden layers that are LSTM layers. Then, we add a dense layer having the ReLU activation function. In order to avoid overfitting, we use a Dropout with $\alpha = 30$. We finally used a fully connected layer, with the softmax as an activation function and the same number of classes, 64.

Both CNN and LSTM-RNN models were trained for 100 epochs. The performances of the systems were measured with the accuracy performance metric. We present with Fig. 4 examples of the accuracy curves. We give a first example using 15 s of training and 2 s of test speech duration with the proposed LSTM-RNN based system and the CNN based system using CMVNC features with NIST SRE 2010 database and a second example with 15 s of training and 2 s of test speech durations using CMVNC coefficients with the LSTM-RNN and CNN based systems for VoxCeleb2 database.

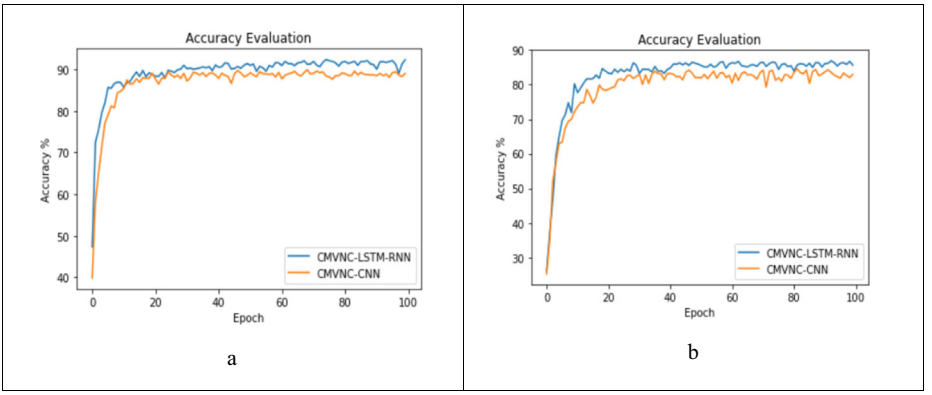

Fig. 4 Speaker recognition accuracy for 15 s of training and 2 s of testing data durations with the proposed CNN based system and LSTM-RNN based system and using CMVNC coefficients for NIST SRE 2010 and VoxCeleb2 databases. **a** Recognition accuracy with NIST SRE 2010 database, **b** recognition accuracy with VoxCeleb2 database

Figure 5 shows and resume the best speaker recognition accuracy obtained from the different set of experiments evaluated with the different systems where 15 s of speech duration per speaker are used in the learning task and speech segments having duration of 2 s, 1 s and 0.5 s per speaker are used in the testing task for the different databases.

Evaluating the performance of the different systems using the standard MFCC features with the two databases show that the CNN based system and the LSTM-RNN approach achieved comparable results with that obtained with the baseline i-vector-PLDA system. In fact, with NIST SRE 2010 database, the best achieved results with 2 s of utterance duration for testing are 84.38% and 75% respectively with the baseline i-vector-PLDA and CNN based systems. The LSTM-RNN system have superior performance since it achieved a recognition accuracy of 88.28%.

The reduction of the speech utterance duration decrease the systems performances which achieved only 82.81%, 71.88% and 85.94% of recognition accuracy respectively with the baseline i-vector-PLDA, CNN based system and the LSTM-RNN system.

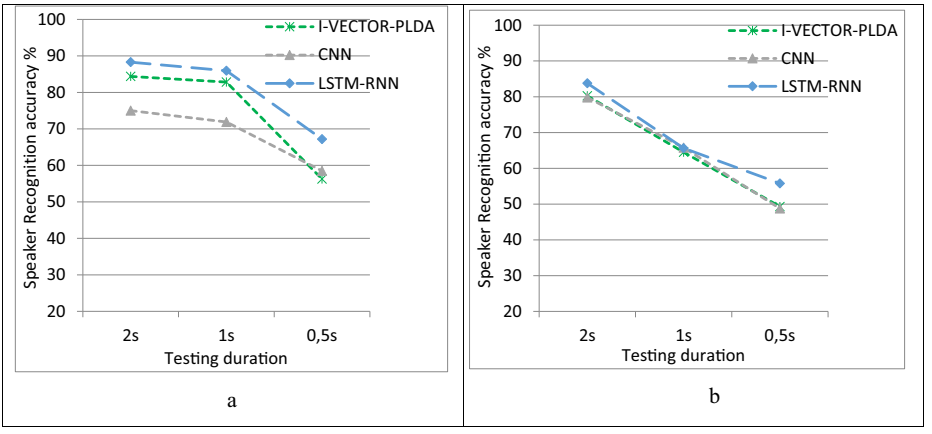

Fig. 5 Speaker recognition accuracy for 15 s of training and different testing durations using MFCC coefficients for NIST SRE 2010 and VoxCeleb2 databases. **a** Recognition accuracy with NIST SRE 2010 database, **b** recognition accuracy with VoxCeleb2 database

The use of very short test speech utterance of 0.5 s further decrease the systems performances which achieved only 58.49% and 67.16% respectively with the baseline CNN based system and the LSTM-RNN systems. In this case, the i-vector-PLDA achieved inferior recognition performance and achieved 56.25% of recognition accuracy.

Similar remarks can also be deduced with experiments evaluated on the VoxCeleb2 dataset. Indeed, the speaker recognition performance with 2 s of test utterance duration is about 80.25% with the i-vector-PLDA system. It slightly decrease with the CNN based system that achieve 79.81% and it reaches 83.78% with the LSTM-RNN system. The LSTM-RNN system also succeed to ameliorate the speaker recognition performance and achieve 55.78% of recognition performance when only 0.5 s of speech utterance duration are used for testing against only 49.31% and 48.78% of recognition accuracy obtained respectively with the i-vector-PLDA system and the CNN based system.

In order to enhance the systems performance, we normalize the cepstral coefficients and we use the proposed CMVNC coefficients. In fact, these features help to more capture the spectral variations presented in the speech signal.

Although accuracy is the most intuitive performance measure, other performance parameters such precision and recall was computed to judge the performance of this model [24]. Precision is defined as the ratio of true predicted positive to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations, also called the true positives, to all observations in actual class. We present with Tables 1 and 2 examples of the performance parameters calculated For 15 s of training and different durations of test speech utterances with the proposed LSTM-RNN based system using CMVNC features with NIST SRE 2010 database and VoxCeleb2 database.

For the proposed LSTM-RNN based system using CMVNC features with NIST SRE 2010 database, the best accuracy of 92.19% has been obtained for 2 s of test speech utterance duration. Precision of 90.89% indicates that the model has accurately predicted positives 90% of times. A recall of 92.15% achieved is of a good performance as it's above 50%.

The following Fig. 6 explain more the difference in terms of speaker recognition accuracies for the different set of experiments.

From these results, we can notice the benefit of the employment of the proposed CMVNC features with the different systems. In fact, the best-achieved result with NIST SRE 2010 database is 92.19% with the proposed LSTM-RNN system for 2 s of test speech data, however it is only 88.28% and 89.06% % respectively with the baseline i-vector-PLDA system and the CNN based system. With the reduction of test speech duration to 0.5 s, the performances of both CNN based system and the LSTM-RNN system achieved respectively 66.44% and 70.79% which outperform the i-vector-PLDA system that realize only 59.81% of speaker accuracy.

At this stage, if we compare between the proposed system in this work and the systems evaluated in [8], taking the nearest condition to our present work, we can cite for example that

Table 1 Speaker recognition accuracy for 15 s of training and different testing durations with the proposed LSTM-RNN based system using CMVNC coefficients for NIST SRE 2010 database

Test duration	Precision	Recall	Accuracy
2 s	90,89	92,15	92,19
1 s	84,82	84,98	85,94
0,5 s	69,87	70,65	70,79

Table 2 Speaker recognition accuracy for 15 s of training and different testing durations with the proposed LSTM-RNN based system using CMVNC coefficients for VoxCeleb2 database

Test duration	Precision	Recall	Accuracy
2 s	85,32	85,62	85,78
1 s	72,16	75,49	75,63
0,5 s	58,36	60,21	60,25

for 10 s per speaker for the training task and 2 s per speaker for the test task, with NIST SRE 2010 database, the baseline system using the i-vector-PLDA algorithm achieve more than 80% and the proposed system achieve more than 90% which is comparable to results obtained here for 15 s of training and 2 s for testing which are 84.38% with the i-vector-PLDA using the standard MFCC features and 92.19% with the proposed LSTM-RNN approach with the proposed CMVNC features. So, we can deduce that our proposed system is competent and well-functioning. It favourite then the task of short utterance speaker recognition.

The proposed LSTM-RNN using the proposed CMVNC system demonstrate also its effectiveness with VoxCeleb2 dataset. In fact, the speaker recognition performance with 0.5 s of test utterance duration is about 59.88% with the i-vector-PLDA system and only 56.5% with the CNN based system and it reaches 60.25% with the LSTM-RNN system.

4.2 Speaker recognition with 5 s of training data duration

The aim of this section is to examine the speaker recognition performance of the different baseline and proposed systems with limited training data duration. Further experiments are hence carried out with only 5 s of training and the same data duration set up previously adopted for the test task.

Figure 7 presents the obtained speaker recognition accuracy from the different experiments evaluated with the different systems where 5 s of speech utterance duration per speaker are used in the learning phase and speech segments having length of 2 s, 1 s and 0.5 s per speaker are used in the test phase for the different databases.

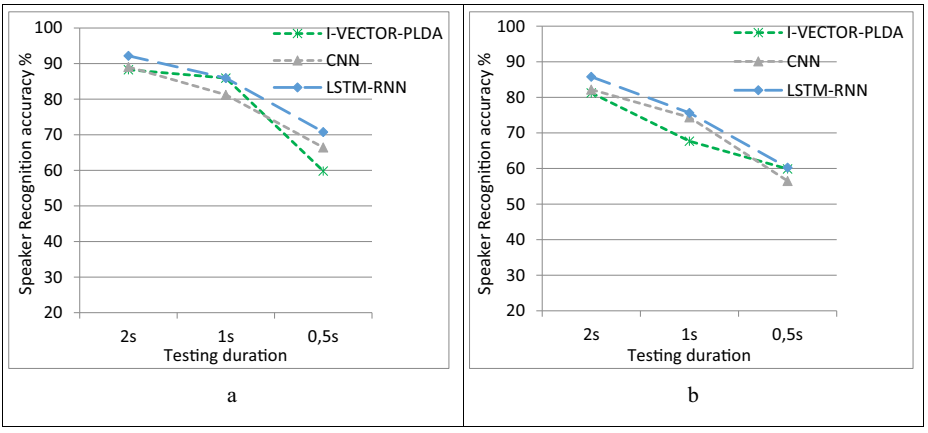

Fig. 6 Speaker recognition accuracy for 15 s of training and different testing durations using CMVNC coefficients for NIST SRE 2010 and VoxCeleb2 databases. **a** Recognition accuracy with NIST SRE 2010 database, **b** recognition accuracy with VoxCeleb2 database

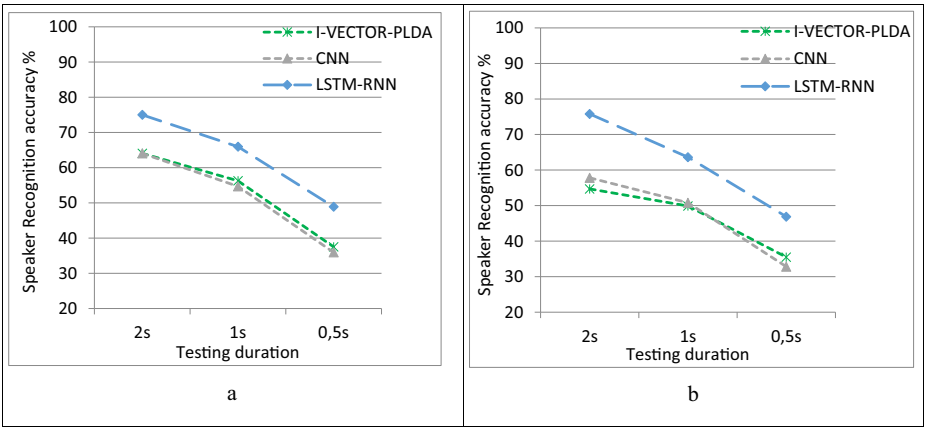

Fig. 7 Speaker recognition accuracy for 5 s of training and different testing durations using MFCC coefficients for NIST SRE 2010 and VoxCeleb2 databases. **a** Recognition accuracy with NIST SRE 2010 database, **b** recognition accuracy with VoxCeleb2 database

As a general observation, we can remark that the speaker recognition performance decrease significantly with the reduction of training data duration. The speaker recognition accuracy achieved with the MFCC features for the baseline i-vector-PLDA system, the LSTM-RNN system and CNN based system are nearly the same. In fact, with NIST SRE 2010 database, for 2 s of testing data duration, the baseline i-vector-PLDA system performance and the CNN based system are both 64.06% and superior performance of 75% is achieved with the LSTM-RNN system. With 1 s of testing data duration, the i-vector-PLDA system performance is 56.25% against only 54.69% with the CNN based system. The use of the LSTM-RNN system increase the system performance which reach 65.94% of speaker recognition accuracy.

With Voxceleb2 database, we can also remark the performance degradation with the limitation of the amount of speech data in the training phase. We can also notice that the LSTM-RNN system achieves superior recognition rates for the most cases when the amount of testing data are varied. Indeed, this system achieves 75.78% of speaker identification rate with

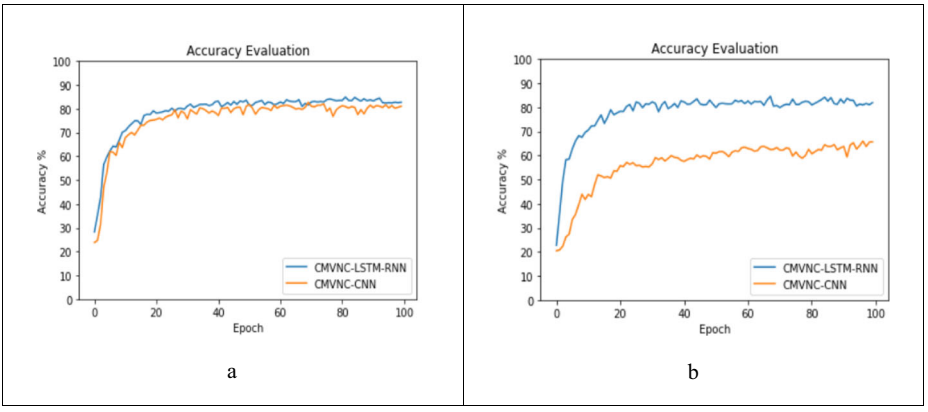

Fig. 8 Speaker recognition accuracy for 5 s of training and 2 s of testing data durations with the proposed CNN based system and LSTM-RNN based system and using CMVNC coefficients for NIST SRE 2010 and VoxCeleb2 databases. **a** Recognition accuracy with NIST SRE 2010 database, **b** recognition accuracy with VoxCeleb2 database

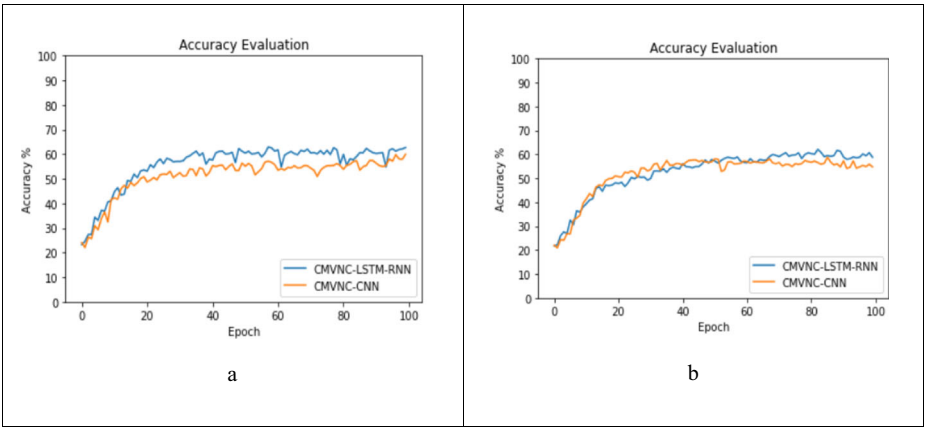

Fig. 9 Speaker recognition accuracy for 5 s of training and 0.5 s of testing data durations with the proposed CNN based system and LSTM-RNN based system and using CMVNC coefficients for NIST SRE 2010 and VoxCeleb2 databases. **a** Recognition accuracy with NIST SRE 2010 database, **b** recognition accuracy with VoxCeleb2 database

2 s of testing against only 54.69% and 57.81% obtained respectively with the i-vector-PLDA system and the CNN based system.

With 0.5 s of testing data duration, the use i-vector-PLDA system permit to obtain a performance of 35.5% against only 32.81% with the CNN based system. In applying the LSTM-RNN system, we notice some improvement in the system's performance which reach 46.88%.

We normalize the cepstral coefficients using the proposed CMVNC coefficients. We present with Figs. 8 and 9 some accuracy curves presenting the performance of the systems with 5 s of training and 2 s of test speech duration with the proposed LSTM-RNN based system and the CNN based system using CMVNC features with NIST SRE 2010 database and VoxCeleb2 database. We give further examples with 5 s of training and 0.5 s of test speech

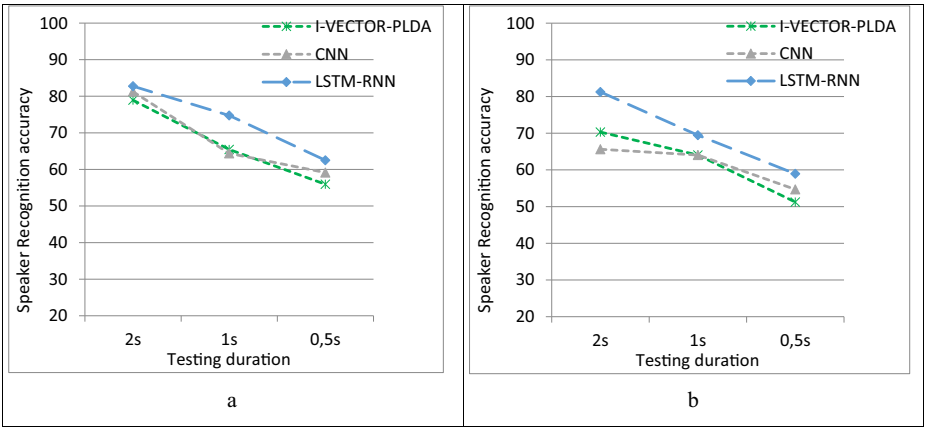

Fig. 10 Speaker recognition accuracy for 5 s of training and different testing durations using CMVNC coefficients for NIST SRE 2010 and VoxCeleb2 databases. **a** Recognition accuracy with NIST SRE 2010 database, **b** recognition accuracy with VoxCeleb2 database

testing durations using CMVNC coefficients with the LSTM-RNN and CNN based systems for both datasets.

The best speaker recognition accuracy obtained from the different set of experiments evaluated with the different systems for 5 s of speech duration are used per speaker in the training task and speech segments having duration of 2 s, 1 s and 0.5 s per speaker are used in the testing task for the different databases are given with Fig. 10.

From the different set of experiments, we can also remark that inferior performances are observed when speech data duration decreases. In spite of this reduction, in applying the CMVNC coefficients, the performance of the proposed LSTM-RNN system using the proposed CMVNC coefficients is significant compared to i-vector-PLDA system and CNN based system and succeeds to achieve 82.75% of recognition accuracy against only 78.94% and 81.25% with 2 s of test utterance durations respectively with i-vector-PLDA system and CNN based system with the NIST SRE 2010 database.

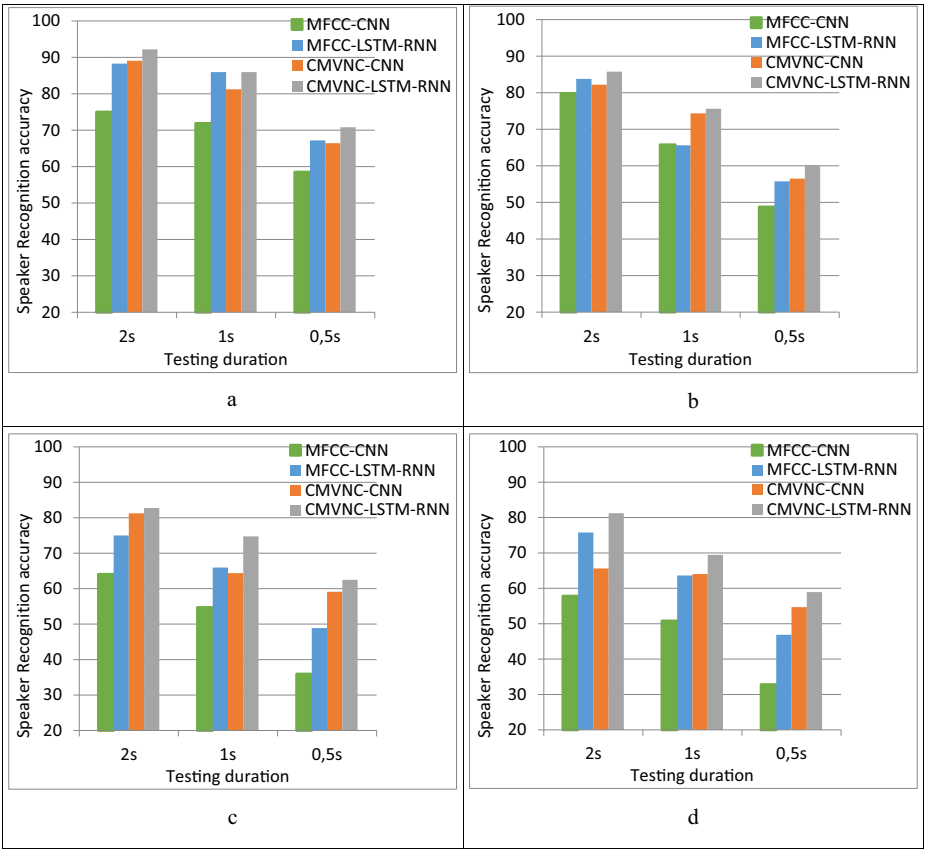

Fig. 11 Speaker recognition performance with the CNN based system and LSTM-RNN system using MFCC coefficients and the proposed LSTM-RNN system and CNN based system using the proposed CMVNC coefficients for 15 s and 5 s of training and different testing durations for NIST SRE 2010 and VoxCeleb2 databases. **a** Recognition accuracy for 15 s of training with NIST SRE 2010 database, **b** recognition accuracy for 15 s of training with VoxCeleb2 database. **c** Recognition accuracy for 5 s of training with NIST SRE 2010 database, **d** recognition accuracy for 5 s of training with VoxCeleb2 database

The performance of the proposed LSTM-RNN system using the proposed CMVNC coefficients fall down to 74.75% with 1 s of testing and 62.5% with 0.5 s of testing. Even though, this performance almost exceeds both performances obtained with CNN based system and i-vector-PLDA baseline systems that gives respectively 65.45% and 64.38% of speaker recognition accuracy with 1 s of testing and only 55.94% and 59.14% with 0.5 s of test speech data duration. Similar remarks can also be deduced with VoxCeleb2 database where the proposed LSTM-RNN system using the proposed CMVNC coefficients succeeds for example to achieve 81.25% of recognition accuracy against only 70.31% and 65.63% with 2 s of test utterance durations respectively with i-vector-PLDA system and CNN based system.

The evaluation of the different systems for most cases let us deduce that the proposed LSTM-RNN system using the proposed CMVNC coefficients give significant improvement compared to the baseline i-vector-PLDA system using the MFCC coefficients. Furthermore, it give superior performance compared to CNN based system using MFCC coefficients and LSTM-RNN system using MFCC coefficients for both significant and reduced training data duration. In order to more clarify the contribution of the proposed systems, we recur to the following Fig. 11 that compare and resume the contribution of the proposed LSTM-RNN system using the proposed CMVNC coefficients and CNN based system using the proposed CMVNC coefficients against the CNN based system using MFCC coefficients and even LSTM-RNN system using MFCC coefficients, in terms of speaker recognition accuracy for different set of experiments evaluated on NIST SRE 2010 and VoxCeleb2 databases.

From the previously obtained results, it can be observed that the use of the proposed approach based on the proposed LSTM-RNN system using the proposed CMVNC coefficients represents an important solution to improve the performance of the speaker recognition systems. This new approach facilitates the task of recognition of the speakers and represent a key element that enhance the performance of systems dealing with reduced speech data duration in both training and testing phases.

5 Conclusions

In this paper, we presented and evaluated a new speaker recognition system based on deep learning approach. We propose a first deep neural network structure with CNN using the proposed CMVNC coefficients and another proposed system based on a LSTM-RNN approach, employing the CMVNC coefficients. The performances exceed that of baseline i-vector-PLDA system using the MFCC coefficients and even CNN and LSTM-RNN approaches with the MFCC coefficients. We then evaluated the systems for different training and testing data duration for NIST SRE 2010 and VoxCeleb2 databases. The novel approach yielded to performance improvement and increase the recognition results, which demonstrates its effectiveness when dealing with short utterance condition challenge. So we succeed to alleviate the effect of short utterance on speaker recognition with the proposed approach. Even though, the system still present some limitations as the utterance durations got shorter. In fact, these results might be further enhanced especially for very short utterance durations. That's why our future research will focus on incorporating additional modelling technique together with the proposed system. In future work, we therefore want to develop more effective strategies for recognizing the speakers in more challenging situations. We will also concentrate on incorporating face detection technique together with signal modality to give a possible more benefit from their joint manipulation.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abd El-Moneim S, Nassar MA, Dessouky MI, Ismail NA, El-Fishawy AS, Abd El-Samie FE (2020) Text-independent speaker recognition using LSTM-RNN and speech enhancement. Multimed Tools Appl 79(33):24013–24028
2. Al-Karawi KA, Mohammed DY (2021) Improving short utterance speaker verification by combining MFCC and Entrocy in Noisy conditions. Multimed Tools Appl 80(14):22231–22249
3. Alam MJ, Kenny P, Stafylakis T (2015) Combining amplitude and phase-based features for speaker verification with short duration utterances. Proc. INTERSPEECH, pp 249–253
4. Bahmaninezhad F, Zhang C, Hansen JH (2021) An investigation of domain adaptation in speaker embedding space for speaker recognition. Speech Comm 129:7–16
5. Bai Z, Zhang XL (2021) Speaker recognition based on deep learning: an overview. Neural Netw 140:65–99
6. Bhattacharya G, Alam J, Kenny P, Gupta V (2016) Modelling speaker and channel variability using deep neural networks for robust speaker verification. In: Proceedings of the 2016 IEEE spoken language technology workshop, SLT 2016, San Diego, CA, USA, December 13-16, pp 192–198
7. Chakroun R, Frikha M (2018) New approach for short utterance speaker identification. IET Signal Processing 12(7):873–880
8. Chakroun R, Frikha M (2020) Robust features for text-independent speaker recognition with short utterances. Neural Comput & Applic 32(17):13863–13883
9. Chakroun R, Frikha M (2020) Efficient text-independent speaker recognition with short utterances in both clean and uncontrolled environments. Multimed Tools Appl 79(29):21279–21298
10. Chiu CC, Lawson D, Luo Y, Tucker G, Swersky K, Sutskever I, Jaitly N (2017) An online sequence-to-sequence model for noisy speech recognition, arXiv preprint arXiv:1706.06428
11. Chung JS, Nagrani A, Zisserman A (2018) Voxceleb2: deep speaker recognition. arXiv preprint arXiv: 1806.05622
12. Dahl GE, Yu D, Deng L, Acero A (2012) Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEETrans Audio Speech Lang Process 20(1):30–42. https://doi.org/10.1109/TASL.2011.2134090
13. Das RK, Prasanna SM (2018) Speaker verification from short utterance perspective: a review. IETE Tech Rev 35(6):599–617
14. Dehak N, Kenny P, Dehak R, Glembek O, Dumouchel P, Burget L, Hubeika V, Castaldo F (2009) Support vector machines and joint factor analysis for speaker verification. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Pro-cessing (ICASSP'09), pp 4237–4240
15. Dehak N, Kenny PJ, Dehak R, Dumouchel P, Ouellet P (2011) Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process 19(4):788–798
16. Devi KJ, Thongam K (2020) Automatic speaker recognition from speech signal using bidirectional long-short-term memory recurrent neural network. Comput Intell
17. Ding I Jr, Ou DC (2015) Enhancements of SVM speaker recognition by dynamic time wrapping. In: Applied mechanics and materials, vol 764. Trans Tech Publications Ltd, pp 891–894
18. Drozdowski P, Rathgeb C, Busch C (2019) Computational workload in biometric identification systems: an overview. IET Biom 8(6):351–368
19. Dua M, Jain C, Kumar S (2022) LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems. J Ambient Intell Humaniz Comput 13(4):1985–2000
20. Fatima N, Zheng TF (2012) Short utterance speaker recognition a research agenda. In: 2012 international conference on systems and informatics (ICSAI2012). IEEE, pp 1746–1750
21. Fei Z, Zhang J-S Softmax discriminant classifier. In: Proceedings of the 2011 third international conference on multimedia information networking and security, Shanghai, China, 4–6 November 2011, pp 16–19
22. Gelly G, Gauvain J-L, Le VB, Messaoudi A A divide-and-conquer approach for language identification based on recurrent neural networks. In: Proceedings of the INTERSPEECH, San Francisco, CA, USA, 8–12 September 2016, pp 3231–3235

23. Ghahabi O, Hernando J (2014) Deep belief networks for i-vector based speaker recognition. In: Proceedings of the 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1700–1704. https://doi.org/10.1109/ICASSP.2014.6853888
24. Ghosh S, Rana A, Kansal V (2019) A statistical comparison for evaluating the effectiveness of linear and nonlinear manifold detection techniques for software defect prediction. Int J Adv Intell Paradig 12(3–4): 370–391
25. Glorot X, Bordes A, Bengio Y Deep sparse rectifier neural networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011, vol 15, pp 315–323
26. Guo G, Zhang N (2019) A survey on deep learning based face recognition. Comput Vis Image Underst 189: 102805
27. Hajavi A, Etemad A (2019). A deep neural network for short-segment speaker recognition. arXiv preprint arXiv:1907.10420
28. Hatch AO, Kajarekar SS, Stolcke A (2006) Within-class covariance nor-malization for SVM-based speaker recognition. In: Proc. Interspeech, Pittsburgh, PA, pp 1471–1474
29. Ho T, Thanh TD (2021) Discovering community interests approach to topic model with time factor and clustering methods. J Inf Process Syst 17(1):163–177
30. Hochreiter S, Schmidhuber J (November 1997) Long short-term memory. Neural Comput 9(8):1735–1780
31. Hong Q, Li L, Li M et al (2015) Modified-prior PLDA and score calibration for duration mismatch compensation in speaker recognition system. Proc. INTERSPEECH, pp 1037–1041
32. Huh JH, Seo YS (2019) Understanding edge computing: engineering evolution with artificial intelligence. IEEE Access 7:164229–164245
33. Ioffe S, Szegedy C Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd international conference on machine learning, Lille, France, 7–9 July 2015, pp 448–456
34. Jansen W (2004) Authenticating mobile device users through image selection. WIT Trans Inf Commun Technol 30
35. Jati A, Georgiou P (2018) An unsupervised neural predictionframework for learning speaker embeddings using recurrentneural networks. INTERSPEECH, pp 1131–1135
36. Jayanna HS, Mahadeva SR (2009) Multiple frame size and rate analysis for speaker recognition under limited data condition. IET Signal Process 3(3):189–204
37. Jia Y, Chen X, Yu J, Wang L, Xu Y, Liu S, Wang Y (2021) Speaker recognition based on characteristic spectrograms and an improved self-organizing feature map neural network. Complex Intell Syst 7(4):1749–1757
38. Kabir MM, Mridha MF, Shin J, Jahan I, Ohi AQ (2021) A survey of speaker recognition: fundamental theories, recognition methods and opportunities. IEEE Access
39. Kanagasundaram A, Dean D, Sridharan S (2014) Improving PLDA speaker verification with limited development data. Proc. ICASSP, pp 1665–1669
40. Kanagasundaram A, Dean D, Sridharan S (2014) Improving PLDA speaker verification with limited development data. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing
41. Kanagasundaram A, Dean D, Sridharan S, Fookes C (2016) Dnn based speaker recognition on short utterances. arXiv preprint arXiv:1610.03190
42. Kanagasundaram A, Dean D, Sridharan S, Ghaemmaghami H, Fookes C (2017) A study on the effects of using short utterance length development data in the design of GPLDA speaker verification systems. Int J Speech Technol 20(2):247–259
43. Kanagasundaram A, Sridharan S, Ganapathy S, Singh P, Fookes C (2019) A study of x-vector based speaker recognition on short utterances. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association, INTERSPEECH 2019. Vol. 2019-September. ISCA (International Speech Communication Association), pp 2943–2947
44. Khosravani A, Homayounpour MM (2018) Nonparametrically trained PLDA for short duration i-vector speaker verification. Comput Speech Lang 52:105–122
45. Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: from features to supervectors. Speech Comm 52(1):12–40
46. Krishnamoorthy P, Jayanna HS, Prasanna SM (2011) Speaker recognition under limited data condition by noise addition. Expert Syst Appl 38(10):13487–13490
47. Kumar GS, Raju KP, CPVNJ MR, Satheesh P (2010) Speaker recognition using GMM. Int J Eng Sci Technol 2(6):2428–2436
48. Laskar MA, Laskar RH (2021) HiLAM-aligned kernel discriminant analysis for text-dependent speaker verification. Expert Syst Appl 182:115281

49. Laskar MA, Bhanja CC, Laskar RH (2021) Speaker-phrase-specific adaptation of PLDA model for improved performance in text-dependent speaker verification. Circ Syst Signal Process 40(10):5127–5151
50. Lei Y, Scheffer N, Ferrer L, McLaren M (2014) A novel scheme for speaker recognition using a phonetically-aware deep neural network. In:Proceedings of the 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1695–1699. https://doi.org/10.1109/ICASSP.2014.6853887
51. Li KP, Wrench EH Jr (1982) Text-independent speaker recognition with short utterances. J Acoust Soc Am 72(S1):S29–S30
52. Li ZY, Zhang WQ, Liu J (2015) Multi-resolution time frequency feature and complementary combination for short utterance speaker recognition. Multimed Tools Appl 74(3):937–953
53. Li L, Wang D, Zhang C, Zheng TF (2016) Improving short utterance speaker recognition by modeling speech unit classes. IEEE/ACM Trans Audio Speech Lang Process 24(6):1129–1139
54. Li D, Liu J, Wang Z, Li Y, Chen B, Cai L (2022) TRSD: a time-varying and region-changed speech database for speaker recognition. Circ Syst Signal Process 41(7):3931–3956
55. Liu Z, Wu Z, Li T, Li J, Shen C (2018) GMM and CNN hybrid method for short utterance speaker recognition. IEEE Trans Industr Inform 14(7):3244–3252
56. Lozano-Diez A, Silnova A, Matejka P, Glembek O, Plchot O, Pesan J, Burget L, Gonzalez-Rodriguez J (2016) Analysis and optimization of bottleneck features for speaker recognition. In: Proceedings of odyssey 2016. International Speech Communication Association, pp 352–357
57. Lu WK, Zhang Q (2009) Deconvolutive short-time Fourier transform spectrogram. IEEE Signal Process Lett 16(7):576–579
58. Mak M-W, Hsiao R, Mak B (2006) A comparison of various adaptation methods for speaker verification with limited enrollment data. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), vol 1, p I–I
59. Marr D (1977) Artificial intelligence—a personal view. Artif Intell 9(1):37–48
60. Matsui T, Furui S (1994) Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's. IEEE Trans Speech Audio Process 2(3):456–459
61. Meghanani A, Anoop CS, Ramakrishnan AG (2021) An exploration of log-mel spectrogram and MFCC features for Alzheimer's dementia recognition from spontaneous speech. In: 2021 IEEE spoken language technology workshop (SLT). IEEE, pp 670–677
62. Nagrani A, Chung JS, Zisserman A (2017) VoxCeleb: a large-scale speaker identification dataset. INTERSPEECH, pp 2616–2620
63. Nainan S, Kulkarni V (2020) Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN. Int J Speech Technol:1–14
64. National Institute Of Standards and Technology, NIST (2010) Speaker recognition evaluation plan. Available at http://www.itl.nist.gov/iad/mig/tests/sre/2010/. Accessed 2010
65. Novoselov S, Pekhovsky T, Kudashev O, Mendelev VS, Prudnikov A (2015) Non-linear PLDA for i-vector speaker verification. In: Proceedings of the 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 214–218
66. Pal M, Saha G (2015) On robustness of speech based biometric systems against voice conversion attack. Appl Soft Comput 30:214–228
67. Poddar A, Sahidullah M, Saha G (2017) Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biom 7(2):91–101
68. Ranzato MA, Huang FJ, Boureau YL, LeCun Y (2007) Unsupervised learning of invariant feature hierarchies with appli-cations to object recognition. In: Computer vision and pattern rec-ognition, 2007. CVPR'07. IEEE conference, pp 1–8
69. Rao K, Sak H, Prabhavalkar R (2017) Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In: 2017 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE., pp 193–199
70. Reynolds DA, Campbell WM (2008) Text-independent speaker recognition. In: Springer handbook of speech processing. Springer, Berlin, Heidelberg, pp 763–782
71. Reynolds DA, Rose RC (1995) Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans Speech Audio Process 3(1):72–83
72. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. Digital Signal Process 10(1–3):19–41
73. Rohdin J, Silnova A, Diez M, Plchot O, Matějka P, Burget L (2018) End-to-end DNN based speaker recognition inspired by i-vector and PLDA. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4874–4878
74. Rohdin J, Silnova A, Diez M, Plchot O, Matějka P, Burget L, Glembek O (2020) End-to-end DNN based text-independent speaker recognition for long and short utterances. Comput Speech Lang 59:22–35

75. Sak H, Senior AW, Beaufays F Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv 2014, arXiv:1402.1128
76. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
77. Shaheed K, Mao A, Qureshi I, Kumar M, Abbas Q, Ullah I, Zhang X (2021) A systematic review on physiological-based biometric recognition systems: current and future trends. Arch Comput Methods Eng: 1–44
78. Snyder D, Ghahremani P, Povey D, Garcia-Romero D, Carmiel Y, Khudanpur S (2016) Deep neural network-based speaker embeddings forend-to-end speaker verification. In: Proceedings of the 2016 IEEE spoken language technology workshop (SLT), pp 165–170. https://doi.org/10.1109/SLT.2016.7846260
79. Soldi G, Bozonnet S, Alegre F et al (2014) Short-duration speaker modelling with phone adaptive training. Proc, Odyssey
80. Song Z (2020) English speech recognition based on deep learning with multiple features. Computing 102(3):663–682
81. Togneri R, Pullella D (2011) An overview of speaker identification: accuracy and robustness issues. IEEE Circuits Syst Mag 11(2):23–61
82. Tran DT, Huh JH (2022) Building a model to exploit association rules and analyze purchasing behavior based on rough set theory. J Supercomput 78(8):11051–11091
83. Vogt R, Sridharan S, Mason M (2010) Making confident speaker verification decisions with minimal speech. IEEE Trans Audio Speech Lang Process 18(6):1182–1192
84. Wang J, Wang K-C, Law M, Rudzicz F, Brudno M (2019) Centroid-based deep metric learning for speaker recognition. IEEE International Conference on Acoustics, Speech and SignalProcessing (ICASSP)
85. Xu C, Rao W, Wu J, Li H (2021) Target speaker verification with selective auditory attention for single and multi-talker speech. IEEE/ACM Trans Audio Speech Lang Process 29:2696–2709
86. Yadav S, Rai A (2020) Frequency and temporal convolutional attention for text-independent speaker recognition. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6794–6798
87. Yamada T, Wang L, Kai A (2013) Improvement of distant-talking speaker identification using bottleneck features of DNN. INTERSPEECH, pp 3661–3664
88. Zhang X, Zou X, Sun M, Zheng TF, Jia C, Wang Y (2019) Noise robust speaker recognition based on adaptive frame weighting in GMM for I-vector extraction. IEEE Access