Neural Computing and Applications (2020) 32:13863–13883
https://doi.org/10.1007/s00521-020-04793-y

ORIGINAL ARTICLE

Robust features for text-independent speaker recognition with short utterances

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

Received: 12 July 2019 / Accepted: 17 February 2020 / Published online: 10 March 2020

Abstract
Speaker recognition systems achieve good performance under controlled conditions. However, in real-world conditions, the performance degrades drastically. The principal cause being when limited data are presented. The presence of background noise is another main factor of performance distortion. In spite of the major advances in speaker recognition field, the effect of noise and the limitation of the amount of available speech data are still open problems, and no optimal solution has been found yet to cope with them. In this paper, we propose a new system using new enhanced and reduced gammatone coefficients in order to improve robustness with limited speech data duration. We demonstrate the usefulness of these coefficients compared to the well-known features with speakers taken from different databases recorded under different conditions.

Keywords Speaker recognition · Speaker identification · i-vector · PLDA · Short utterances · Noise

1 Introduction

Speaker recognition is the ability to recognize an individual only from his voice. This domain has received much attention from the scientific community since many years up to the present day [1–3]. In fact, this technique makes possible the use of the speaker's voice to verify the identity of the user and control the access to many services such as voice dialing, telephone shopping, banking by telephone, database access services, voice mail, information services, security control in confidential information areas, and remote access to the computers. In this manner, speaker recognition technology is expected to create new services that will make our daily lives more appropriate.

✉ Rania Chakroun
rania.chakroun@enis.rnu.tn

Mondher Frikha
mondherfrikha05@yahoo.fr

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

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

[3] National School of Engineering of Sfax, Sfax, Tunisia

Speaker recognition is a big area that can be divided into two fundamental applications which are speaker identification and speaker verification. For the identification task, an unknown speaker is compared against a dataset of known speakers, and the best matching speaker is considered as the identification result. For the task of verification, the system purpose is to make a decision whether a voice sample was produced by the claimed person. Both speaker identification and speaker verification applications can be divided into text-dependent and text-independent methods. In text-dependent systems, speaker recognition depends on a specific text being spoken. This method is simpler to the system. For text-independent systems, there are no limitations for the text used in the test or in the train phase and the speaker must be recognized independent of what is saying. This kind of application, of course, is more complex to handle for the system.

During the last years, the interest in speech and speaker recognition applications over fixed telephone, mobile phone, and handheld devices has been augmented. These devices are almost used in adverse environments such as city streets, airports, offices, and cars. The use of these different means is constantly increasing among the private users and business customers. In addition to the environmental noise in which the speech was produced, telephone

communication channels introduce additional distortions to the speech. These different noise sources alter the speech production so that most of speaker recognition systems are vulnerable to failure in noise corrupted environments.

In recent years, consistent research has been made to cope with the degradations introduced by the presence of background noise on speaker recognition systems. However, no satisfying solution has been found yet [4–6].

The development of a speaker recognition system cannot be complete without taking into account of both the effects of the real-life environment and the requirement of realistic applications. For the real-life environment, the speaker recognition systems must deal with techniques able to combat the degradations introduced by noisy conditions. Concerning the requirement of realistic applications, the system should take into consideration the problems related to the memory and computational resource limitation. For that, the system should be performed with the simplified as possible of algorithm and the minimum as possible of speech utterance durations. In this context, the short utterance speaker identification is so required to develop a performing realistic application. In fact, to ensure proper access to confidential information, personal transactions, and security-related applications, in a realistic application, there are many circumstances and constraints related to the limitation of computing resources, the conditions in which the speech was collected, that impose the reduction in the amount of speech data. Conventionally, the performance of the state-of-the-art speaker recognition systems is very good when sufficient speech data are available for training and testing. It refers to the case in which the system used few minutes (> 1 min) of speech data segments [7, 8] which permit to provide enough feature vectors to fulfill the feature space and then to form well-trained models [9]. By the same token, sufficient data warrants reliable decision for testing. Even so, the performances of the speaker recognition applications have been usually substantially degraded when only limited data are available. This refers to the employment of few seconds (< 15 s) of speech data for the training and testing tasks [7–9]. In this case, there are less feature vectors for training and testing, and hence we have poor modeling and unreliable decision for testing. For that, different methods started to develop in order to address the research problems of short utterance speaker recognition (SUSR), which is now becoming a major consideration of modern speaker recognition research [9–13].

A deep look into speaker recognition domain, and a special concern on short utterance speaker recognition, lets us to deduce that the most commonly used state-of-the-art speaker recognition algorithms address the problem of speaker verification when speech duration is short [13–18]. Since the lack of a particular focus on the problem related to speaker identification based on short utterances, we want with this study to pay a specific attention to speaker identification when a little amount of speech is available. In fact, we will focus on a speaker identification system taking into account that there is no restriction regarding the text content of the input speech utterance, only a little amount of speech is available for training and testing and we will consider also the presence of the background noise. In this context, we examine the effect of speech utterance duration on the system performance, and we try to solve the problem of noisy short utterances through different processes. Indeed, we propose a novel speaker identification system based on new reduced features detected from the speech signal which are the RMNGFCC (Reduced Mean Normalized Gammatone Frequency Cepstral Coefficients) and RMVNGFCC features (Reduced Mean and Variance Normalized Gammatone Frequency Cepstral Coefficients). These features take advantage of feature normalization process like Cepstral Mean Normalization (CMN) and Cepstral Mean and Variance Normalization (CMVN) which help to reinforce the speaker characterization and improve robustness when the used utterances have a noisy limited duration. We show that the use of the proposed features facilitates more the detection of the identity of a person even when there are noisy conditions. The fusion of both RMNGFCC and RMVNGFCC features further improve the robustness of the proposed speaker identification system. A comparison with state-of-the-art systems using state-of-the-art features and standard modeling techniques is presented to highlight the contribution of the proposed approach.

The rest of this paper is organized as follows. Previous works on short utterance and noisy speaker recognition are given in Sect. 2. The proposed speaker identification system is presented in Sect. 3. Experimental results are discussed in Sect. 4, and we draw our conclusions in Sect. 5.

2 State-of-the-art speaker recognition techniques

During the last few years, the speaker recognition field has gained great popularity in a wide range of applications such as speech communications, access control, forensic evidence provision, domestic services, and smart terminals. Current speaker recognition systems have achieved satisfactory performance, given that the enrolment and testing utterances are sufficiently long and the signal is recorded under acceptable conditions [19].

Most state-of-the-art speaker recognition engines referred to generative models like Gaussian mixture models (GMM) to achieve the recognition capability [2]. Indeed,

when speaker models are trained with sufficient amount of data, phonemes are well captured from the speaker, which can lead to better representation of the speaker's acoustic space and help to improve its discriminating ability. This was especially true since the GMM were introduced in order to model the acoustic space for speaker recognition [20]. Hence, the GMM were considered as the most popular tool for state-of-the-art speaker recognition applications [2, 20]. The high success gained by the GMM encourages researchers to look for more improved tools. Thus, the occurrence of effective approaches like the clustering technique is considered recently as an essential equivalently to the GMM approach [21–23].

However, in realistic applications, it is very common that enough data may not be available for speaker training and the test utterances are very short during the recognition task. For such conditions, the GMM is failing to recognize a short utterance speaker with a high accuracy. Indeed, the attempt of using smaller amount of data leads to great performance degradation when dealing with a speaker recognition system. That's why the subsequent research endeavors focused on developing GMM with new techniques such as the i-vector [24–28]-based speaker recognition systems. In spite of that, these applications are still tending to performance degradation when short-duration utterances are used for speaker recognition [25].

The effect of short utterance duration is considered as one of the recent challenges in speaker recognition that was organized by the National Institute of Standards and Technology (NIST) [29], which led the research community to further concentrate on this problem. To overcome this difficulty, a considerable amount of study is going on in order to develop suitable methods when either the given speech is too small or with the aim of using fewer amount of speech to cut computation costs.

Among the earlier works, we notice the use of new classifiers like the SVM for short utterance speaker recognition [9, 10, 30–34]. We found also that the problem was addressed to the use of the deep learning technique [11]. In fact, convolutional neural network (CNN) [11], recurrent neural networks (RNN) [35], and deep neural networks (DNN) [36–38] have been used for speaker recognition systems when using small speech utterances. However, most these works have targeted text-dependent speaker verification [11, 36–38]. More recent works demonstrate that the use of deep network provides better performance for short-duration text-independent speaker verification systems [35, 39–41]. Even so, to the best of our knowledge, deep learning technique been applied to related problems such as speaker verification, and there is a lack of effective recognition method for the short utterance text-independent speaker identification task.

Speaker recognition needs a large amount of speech data, leading to the use of huge files and complicated processing. This has encumbered the speaker recognition technology to be used widely. Researchers have thus led to incorporate new techniques to improve baseline approaches like the Probabilistic Linear Discriminant Analysis (PLDA) approach which is used to improve the i-vector model with short utterance speaker recognition [42, 43]. This combined approach has become dominant and demonstrates to be lately efficient and successful [9, 13, 44].

As one common case of robust speech processing, recognizing speakers from short utterances contaminated by noise is a rather challenging task that has been of interest in several recent studies [45–48]. Hence, researches try to adopt some enhanced approaches or features [49, 50]. Using denoising techniques proved to be essential in handling noise [51]. Feature level enhancement using uncertainty-of-observation techniques [48, 52], vector Taylor series [53] or Higher-Lag Autocorrelation Coefficients [54] helps in robust speaker modeling and recognition. Then, the use of Gammatone filters has gained popularity in several branches of signal processing, including robust speech recognition [55] and speaker recognition [47, 56–58].

Together with the advancement of modern technologies, various methods started to develop for Automatic Speaker Recognition. In spite of the realization of high outperforming algorithms in this domain, these systems are prone to have performance degradation when short utterances are met in the enrolment and test phases. What is more appreciating is that the achieved results still depend on the duration of the speech used for both training and testing task, the questioned task (speaker verification and speaker identification), the dependency to the context (text-dependent, text-independent), the number of the speakers used, the features used, their dimension and the different parameters of the employed approach, etc. The presence of background noise which presents alone another main deteriorating factor for speaker recognition applications may further aggravate the situation in this case. Hence, the challenging area of short Utterance speaker recognition remains an open problem and robust handling of real-world data is still a defiant topic. Since there is a lack of efficient methods for the problem of robust short utterance text-independent speaker recognition, our interests in this study concern the improvement of a text-independent speaker recognition application that has a particular focus on short utterance speaker identification task dealing with the presence of background noise. Hence, we choose to benefit from the most latest successful i-vector based on the use of the PLDA technique for the proposed system. To improve

robustness, an appreciated enhancement is assigned to the feature level which takes advantage from the most recent robust Gammatone filters. This kind of application meets the need for realistic applications which are prone to the effects of the real-life environment, the constraints and the requirements of realistic applications.

3 Proposed approach for robust short utterance speaker identification

3.1 Motivation

Speaker recognition is the method used to recognize persons from their voice. This technique attempts to cover the different aspects for speaking. In fact, each speaker has his own manner of speaking, including his particular accent, rhythm, intonation, style, pronunciation, etc. Thus, the system needs to deal with sufficient utterance duration in order to capture the speaker-specific characteristics and to achieve good performance. However, in a real circumstance, it might be difficult to collect a large amount of speech data as required by conventional speaker recognition approaches. For example, there might be some conditions which oblige a person to speak only a little amount of speech like his state of health, his character, etc. In real life, there are many circumstances that permit only to obtain small amount of clear speech. In fact, speech obtained could be broken or unclear, or recorded in noisy situations or contains some breaks and a little amount of real speech. Moreover, 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 [9]. These entire conditions make short utterance speaker recognition arises as an important area of research in such cases. Along with the advancement of speaker recognition technology, the case of using short data duration remains a major problem. In fact, the use of short segments of speech for recognition purpose leads to great system performance degradation. The use speech segments recorded under uncontrolled environment presents another major problem for speaker recognition performance deterioration that received also the attention of the research community.

In this work, we deal with speaker identification purpose for which the main objective is to identify an unknown speaker from a set of registered speakers. An exhaustive survey on the process of speaker recognition, as well as the different methods used for this objective and human voice, let us deduce various relevant factors for the speaker identification process. Among these factors, we notice that the features extracted from the speech signal present a fundamental element for capturing the speaker-specific characteristics which led to differentiate between the different speakers and make a good discrimination.

Many features have been investigated in the literature for speaker recognition purpose [2]. We can cite the Linear Prediction Coefficients (LPCs) [59] which are directly derived from the speaker's speech production model. Perceptual Linear Prediction (PLP) coefficients [60] are also used in this purpose since they are based on human perceptual and auditory processing. Over the last two decades, spectral features have become popular. Researches have shown that these features are successful for different applications based on speech processing such as speech recognition, speech emotion recognition, speaker recognition tasks [61, 62]. The well-known spectral features are called the Mel Frequency Cepstral Coefficients (MFCCs). These features allow obtaining high level of performance due to the use of the perceptually based Mel spaced filter bank processing of the Fourier Transform and the particular robustness to the environment and flexibility that can be achieved using cepstral analysis [2, 34, 63–65]. Recently, the use of Gammatone Frequency Cepstral Coefficients (GFCCs) has gained popularity for robust speaker recognition applications [47, 56–58]. In fact, these features show better performance for the task of speaker recognition with noisy conditions than other features.

In this work, our endeavors are addressed to look for a suitable application for noisy short utterance speaker identification. Hence, we intend to develop a new approach depending on new enhanced features that investigate the robustness of the GFCC features to give more supplementary information that helps to further facilitate the distinction between the speakers when short utterances are used under noisy conditions. This can fundamentally improve the system performance while avoiding the use of additional, lengthy and complicated algorithms requiring more time and memory space which is beneficial especially for real-world applications.

It follows that speaker recognition applications depend on high dimensional feature vectors. However, realistic applications suffer from many constraints related to the memory and computational resource limitation. Thus, we try to examine if the proposed system can perform well with new low dimensional feature vectors able to reduce the memory and time complexity of the system while maintaining good performance on a speaker identification system when short utterances are used under uncontrolled conditions. The results obtained are compared with state-of-the-art applications and evaluations are compared against another existing works.

3.2 Proposed approach

The development of the proposed robust Short Utterance Speaker Identification system was motivated by a desire of obtaining a set of practical features for speaker recognition that are more robust and with the respect to the acoustical variability in their native form, without loss of the system performance when the speech signal have a limited duration, recorded in realistic environment, and with a degree of computational complexity comparable to that of standards MFCC and GFCC coefficients. For that, we choose to develop an approach that provides pragmatic gains in robustness at small computational costs to be more faithful with realistic applications.

To describe the architecture of the proposed speaker identification system, we began by giving the different main blocks that we used for the different baseline and proposed systems in this work. As shown in Fig. 1, the speaker identification system is composed of a succession of different modules to accomplish the learning and the testing phases. During both phases, the feature extraction process is essential to capture the speaker voice characteristics which allow calculating the appropriate i-vectors of the speakers. The use of the PLDA technique is then carried to build an adequate model for each speaker and to recognize the unknown speaker of the test speech utterance after comparing the model of the test speech signal with those of the different speakers constructed in the training phase of the system.

3.2.1 Feature extraction

3.2.1.1 Standard features The most fundamental process commonly applied in all forms of speaker recognition systems is that concerning the extraction of the features vectors from the acoustic speech wave. This process is applied for each frame which can capture the specific characteristics of speakers.

Many features have been investigated for speaker recognition applications where spectral based features have become the most successful and most popular [2]. The well-known are Mel Frequency Cepstral Coefficients (MFCCs) [2, 63]. In this work, MFCC features were extracted with the Hidden Markov Model ToolKit (HTK) [66]. We perform our experiments with cepstral features extracted using a 25-ms Hamming window having 10-ms overlap. The feature vectors are 12 MFCC calculated every 10 ms together with log-energy and augmented with delta and double-delta coefficients giving 39-dimensional feature vectors. Indeed, this feature vector is widely employed in the state-of-the-art applications [2, 8, 67].

To improve robustness, Gammatone Frequency Cepstral Coefficients (GFCCs) are equally extracted from the speech signal. In fact, the GFCC-based speaker identification is found to achieve a very robust performance, as presented in [68, 69]. According to the observation of [69], most information remains in the lower 23-order GFCC coefficients. Since the zeroth cepstral coefficient was more susceptible to contamination of noise, 22-dimensional GFCC features were used in speaker recognition evaluations [68–70]. IN [71], after using the lower 23-order

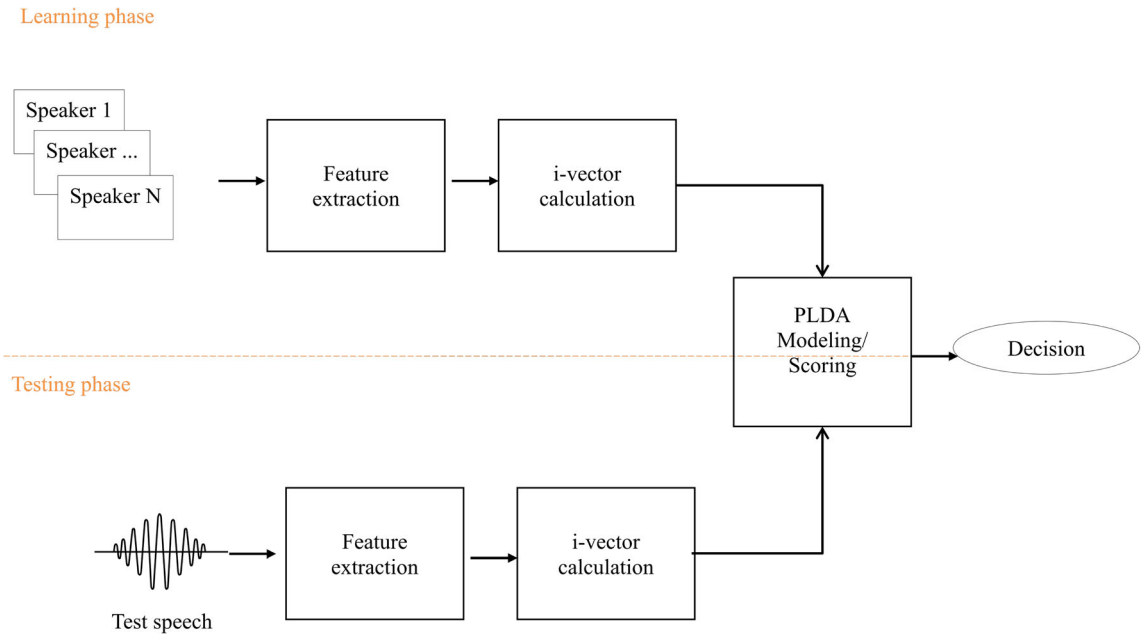

Fig. 1 Automatic Short Utterance Speaker Identification using the proposed system

GFCC as a feature vector in a previous study, the authors find that using 30-dimensional GFCCs as a feature vector is more suitable to retain the information.

In this work, in order to largely retain the information, the first set of experiments with baseline system is dealt with 39-dimensional GFCC feature vectors to represent each frame of the speech signal. The subsequent sets of experiments are dealt with more reduced dimension of GFCC feature vectors in order to find the efficient representation.

3.2.1.2 Proposed features The speaker recognition task supposed to recognize persons from their voice. However, the variability of the speech signal caused by different factors like speaker identity, gender, transmission channel, utterance length, session or speaking style makes this task difficult. It has been proved in the literature that these variations have a direct negative impact in the system performance [20]. That is why compensation techniques at different levels such as feature or score level are needed to cope with speech variability. In this work, in order to diminish the effect of the variability of the extracted features from a session to another, we recur to the Cepstral Mean Normalization (CMN) [2]. In order to more refine the performance and the robustness of the system, we refer to a further normalization which is Cepstral Mean and Variance Normalization (CMVN) [2].

Indeed, in this study, the performance of the proposed method is evaluated with new proposed features and the achieved results were compared to those obtained from two traditional baseline feature-based methods, MFCC and GFCC [68]. We propose new feature vectors in which we use a novel description of features. We use then:

- MNMFCC: Mean Normalized MFCC, which is a short-time cepstral representation of a speech in which we normalize the feature vector coefficients using the CMN technique. In fact, if we note by $X = \{x[n]\}$ where $0 < n \leq N$ the MFCC cepstral vector, then the normalized features presenting the MNMFCC coefficients, presented by $Y = \{y[n]\}$, are calculated as follows:

$$y[n] = x[n] - \frac{1}{N}\sum_{n=1}^{N} x[n] \tag{1}$$

  where N represents the number of MFCC coefficients in a feature vector and n is the order of the MFCC coefficient in this vector.
- MVNMFCC: 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. Indeed, for example, for a given feature vector $X = \{x[1], x[2], \ldots, x[N]\}$ of MFCC coefficients, the vector presenting the MVNMFCC coefficients presented is calculated as follows:

$$\hat{x}[n] = \frac{x[n] - \bar{x}}{\sigma_x} \tag{2}$$

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x[n] \tag{3}$$

$$\sigma_x^2 = \frac{1}{N}\sum_{n=1}^{N} (x[n] - \bar{x})^2 \tag{4}$$

- MNGFCC: Mean Normalized GFCC, which is a short-time cepstral representation of a speech in which we normalize the feature vector coefficients using the CMN technique. In fact, for $Z = \{z[n]\}$ where $0 < n \leq N$ the GFCC cepstral vector, the normalized features presenting the MNGFCC coefficients, presented by $T = \{t[n]\}$, are calculated as follows:

$$t[n] = z[n] - \frac{1}{N}\sum_{n=1}^{N} z[n] \tag{5}$$

  where N represents the number of GFCC coefficients in a feature vector and n is the order of the GFCC coefficient in this vector.
- MVNGFCC: Mean and Variance Normalized GFCC, which is a short-time cepstral representation of a speech in which we normalize the feature vector coefficients using the CMVN technique for that, for a given feature vector $Z = \{z[1], z[2], \ldots, z[N]\}$ of GFCC coefficients, the vector presenting the MVNGFCC coefficients presented is calculated as follows:

$$\hat{z}[n] = \frac{z[n] - \bar{z}}{\sigma_z} \tag{6}$$

$$\bar{z} = \frac{1}{N}\sum_{n=1}^{N} z[n] \tag{7}$$

$$\sigma_z^2 = \frac{1}{N}\sum_{n=1}^{N} (z[n] - \bar{z})^2 \tag{8}$$

The challenge of the speaker recognition task supposed to recognize the identity of the speaker using realistic applications that suffers from computational resource

limitation. That is why, our endeavors are addressed for researching features needing less memory space and then reducing more the memory and time complexity of the system. Several experiments are then done in order to find the efficient representation and let us deduce that the use of the proposed MNGFCC and MVNGFCC features is more efficient with more reduced dimensional feature vectors. For that, the following set of features is equally used in this work:

- RGFCC: Reduced Gammatone Frequency Cepstral Coefficients, which is a short-time cepstral representation of a speech in which we use less coefficients than those used in the standard GFCC feature vector. We note with $Z_r = \{z_r[m]\}$ where $0 < m \leq M$ and $M \leq N$ the RGFCC cepstral vector. M represents the number of RGFCC coefficients in a feature vector and m is the order of the RGFCC coefficient in this vector. N represents the dimension of the standard GFCC feature vector. The chosen dimension of the RGFCC feature vector is explicated in the following Sect. 4.
- RMNGFCC: Reduced MNGFCC, which is a short-time cepstral representation of a speech in which we use lower feature vector dimension than used in the MNGFCC feature vector. We note with $T_r = \{t_r[m]\}$ where $0 < m \leq M$ and $M \leq N$ the RMNGFCC cepstral vector. M represents the number of RMNGFCC coefficients in a feature vector and m is the order of the RMNGFCC coefficient in this vector. N represents the dimension of the MNGFCC feature vector. We refer to Eq. 5 to calculate the feature vector coefficients. The dimension of the RMNGFCC feature vector is explicated in Sect. 4.
- RMVNGFCC: Reduced MVNGFCC, which is a short-time cepstral representation of a speech in which we use lower feature vector dimension than used in the MVNGFCC feature vector. We note with $\hat{z}_r = \{\hat{z}_r[m]\}$ the RMNGFCC cepstral vector where $0 < m \leq M$ and $M \leq N$. M is the dimension of the RMVNGFCC feature vector and m is the order of the RMVNGFCC coefficient in this vector. N represents the dimension of the MVNGFCC feature vector. We refer to Eq. 6 to calculate the feature vector coefficients. The chosen dimension of the RMVNGFCC feature vector is explicated in the following Sect. 4.
- FRMVGFCC: Fused Reduce Mean and Variance Normalized GFCC, which is a short-time cepstral representation of a speech in which we combine both RMNGFCC and RMVNGFCC feature vector coefficients. In fact, the features presenting the FRMVGFCC coefficients, presented by $F = \{F[m]\}$, are calculated as follows:

$$F[m] = [t_r[m]; \hat{z}_r[m]] \tag{9}$$

where $0 < m \leq M$ and $M \leq N$. M is the dimension of the RMVNGFCC and RMNGFCC feature vectors and m is the order of the corresponding coefficient in the appropriate vector. The choice of M is given in Sect. 4. N represents the dimension of the MNGFCC and MVNGFCC feature vectors.

A diagram comparing between the different process of feature extraction for baseline and proposed systems is presented with Fig. 2.

The speaker recognition process comprised two phases which are the learning phase and the testing phase.

3.2.2 Speaker learning

The combination of i-vector [24] and PLDA [72] has become recently a dominant approach for text-independent speaker recognition applications. In fact, the PLDA on the i-vectors has been successfully used and demonstrates to be effective compared to state-of-the-art speaker recognition approaches [9, 13, 67].

In this work, we will model the speaker identification problem with the i-vector-PLDA approach. We have shown that different features are proposed for the speaker recognition purposes. Each proposed feature is evaluated with the i-vector-PLDA approach. As demonstrated in the following section, the experiments performed prove the superiority of both RMNGFCC and RMVNGFCC features with the i-vector-PLDA approach. Below we summarize the algorithm of the proposed robust short utterance speaker identification system training using the i-vector-PLDA technique with the proposed features:

Algorithm 1: speaker learning (SLE)

Input: speech signal belonging to the speaker *l*

Output: speaker model *mod l*

1. Extract RGFCC features
2. Normalize RGFCC features,

   Extract RMNGFCC features
3. Normalize RGFCC features,

   Extract RMVNGFCC features
4. Combining the resulting feature vectors,

   Extract FRMVGFCC features
5. Set as a stop-learning condition (reach the minimum of error or reach max of iteration).
6. Calculate the output of the learning which is the speaker model *mod l*.
7. If the number of iterations or the minimum of error is reached, learning converge, the learning stops; otherwise we return to 5.

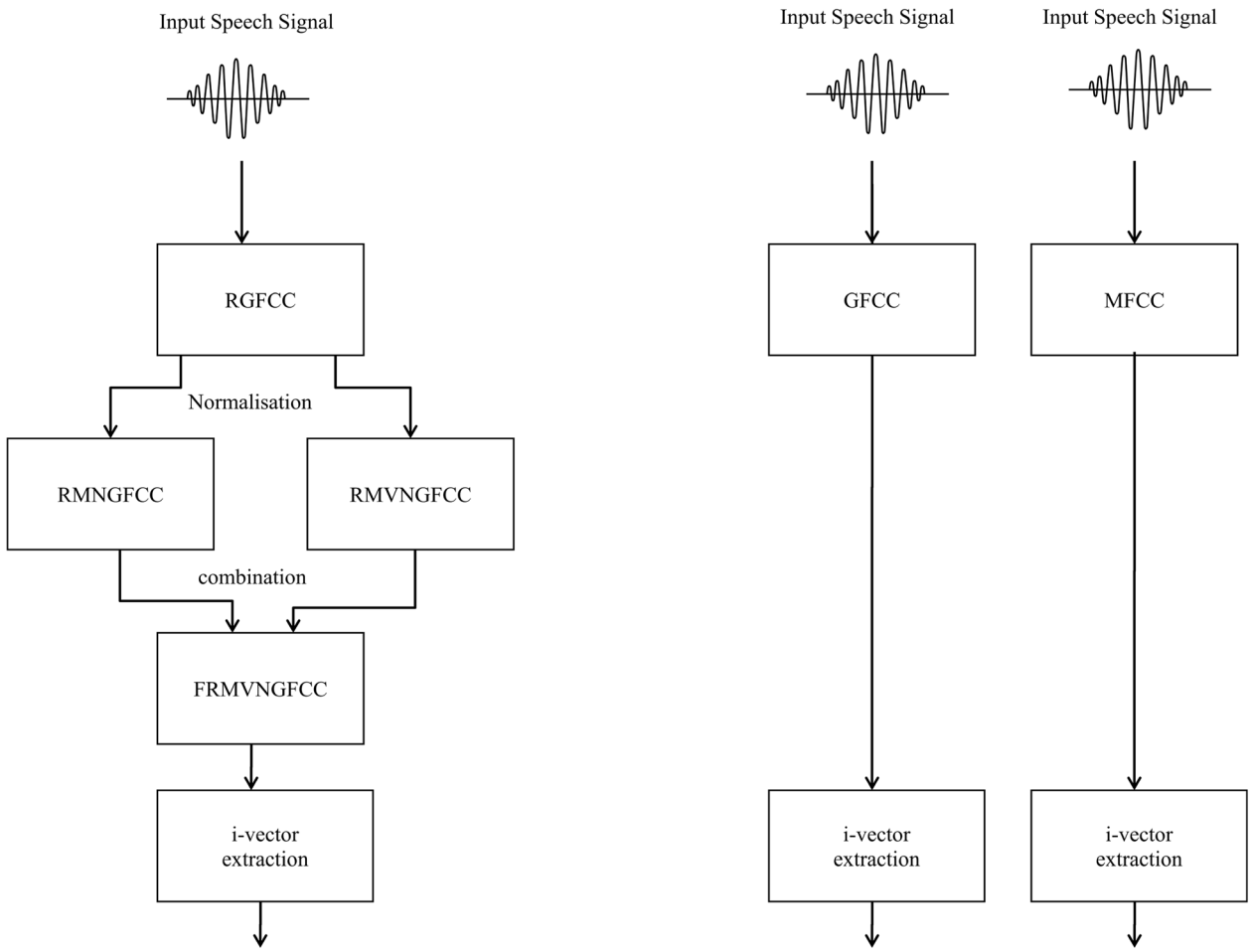

Fig. 2 Comparison of the feature extraction process between MFCC-based baseline system, GFCC-based baseline system and the proposed system using FRMVGFCC features

3.2.3 Speaker recognition

In order to recognize the speaker in the test phase, a test speech utterance is input to the system and the appropriate features are extracted from the speech signal. The adequate model of the test utterance is compared to the models of the different speakers learned with the system with the aim of identifying the most suitable speaker. Hence, the extracted parameters are very essential and represent the requested information needed to facilitate the research of the appropriate speaker. Then, improves the system performance and makes the short utterance sufficient for training and detecting the identity of the speaker. The following algorithm summarizes the speaker identification process.

Algorithm 2: Speaker Identification (SID)

Input: -speech signal belonging to the speaker S

Output: *Identity*(S)

1. Extract RGFCC features
2. Normalize RGFCC features,
   Extract RMNGFCC features
3. Normalize RGFCC features,
   Extract RMVNGFCC features
4. Combining the resulting feature vectors,
   Extract FRMVGFCC features
5. For $j = 1,\ldots, NS$
   Identity (S) ← selecting the most suitable speaker identity among NS speakers

End

Where the abbreviations cited in the above algorithm are the following:

- S present the unknown speaker.
- NS present the number of classes of the different speakers.
- Identity(S) is defined as the most convenient speaker among NS speakers with the appropriate classifier

4 Experiments and discussion

4.1 Corpora

We conducted our experiments with three different databases. The first set of experiments is carried out with the TIMIT Database. This corpus has been primarily designed to provide speech data for the acquisition of acoustic–phonetic knowledge and for the development and evaluation of automatic speech recognition systems [73]. Then, it is widely used in speaker recognition studies [74]. TIMIT database contains 10 different sentences from each of 630 speakers including 438 males and 192 females [75] from eight major dialect regions of the USA. The dataset contains about 5.25 h of audio file in wav format having 16 kHz of sampling frequency with a resolution of 16-bits. The recordings are single channel, and the mean duration of each utterance is 3.28 s.

The second set of experiments is subsequently dealt with the NTIMIT Database [76]. The NTIMIT corpus has been made to enable researchers to perform experiments that compare speech and speaker recognition performance obtained with high-quality speech transmitted over long-distance telephone lines. The NTIMIT database was created by transmitting sentences in the TIMIT database over a physical telephone network. The NTIMIT utterances accurately reflect the general nature of telephone-based speech. The USA is divided into Local Access and Transport Areas (LATAs) which represent the geographical regions corresponding to the subdivision of the telephone network. Within each LATA, various central offices are made to handle calls. Different telephone channels are used to collect the NTIMIT corpus by changing the central office in order to transmit the TIMIT utterances for different geographical locations. In total 253 central offices, and then 253 different telephone channels, are used for the compilation of the NTIMIT database.

Another attempt to study the effectiveness of the proposed system has been carried out using the most recent NIST Speaker Recognition Evaluation data, NIST SRE 2010 corpora [77]. The data had multiple channels, including telephone, microphone, and interview data, as determined from the keys released by NIST [77].

4.2 Experimental protocol

In order to make a comparison with the latest works, experiments were conducted following the protocol suggested in [9] which follows in turn the protocol suggested in [2]. For that, with TIMIT database, the evaluations were performed with 64 speakers from all the 8 regions of the

database. The speakers were selected as 4 male and 4 female speakers from each dialect region.

The second set of experiments is evaluated with the same speakers keeping then the same conditions from the NTIMIT corpora and the third set of experiments is evaluated with the same number of speakers having the same gender conditions from different regions of the USA with the NIST SRE 2010 corpora.

Based on the superior performance of the i-vector based on PLDA (i-vector-PLDA) system for speaker recognition purposes [9, 67], we choose to handle the speaker identification experiments with a baseline system using the i-vector-PLDA technique.

We evaluate the performance of the proposed system against the baseline system for speaker identification task for different training and testing durations by carrying out experimental evaluations as follows.

4.2.1 Speaker identification with an important training duration

To evaluate the systems' performance, the experiments are done with different training and testing duration. At the beginning, we choose to evaluate the system performance with a considerable amount of training data duration (> 15 s) [7–9]. For that, we use approximately 24 s of speech data duration

for the training task and 6 s of speech data duration for the test task. We start by evaluating the performance of the i-vector-PLDA baseline systems with standard MFCC feature vectors for speaker identification purpose. Since we intend to improve more the system performance in such cases, our research endeavors are concentrated on more challenging information extracted from the speech signal. We evaluate then the speaker identification system with the standard GFCC features and the proposed MNMFCC, MVNMFCC, MNGFCC, and MVNGFCC features with 39-dimensional feature vectors. We varied the number of mixture components for the different systems from 1 to 256 mixtures and the correct Identification Rates (IR) obtained from the different experiments using the different feature vectors with speakers taken from TIMIT, NTIMIT, and NIST SRE 2010 databases are, respectively, presented in Fig. 3.

Generally, the proposed systems using the proposed features demonstrated superiority over the i-vector-PLDA baseline systems using standard MFCC and standard GFCC features. In fact, with TIMIT database, the best-achieved performance is 100% from 16 mixtures with both standard MFCC and GFCC features. The use of the proposed features maintains the same performance for the different features.

The experiments made on NTIMIT database prove the usefulness of the proposed features. In fact, the use of the baseline system with standard MFCC features achieved the best performance of 97.66% of IR. The same performance is attained with standard GFCC features. This performance is enhanced with the use of MNMFCC features and attains 99.22% of IR and the best performance of 100% is achieved with the use of MVNMFCC features. The use of MNGFCC and MVNGFCC features is also beneficial since we can achieve, respectively, 100% and 99.22% of IR.

For the speakers taken from NIST SRE 2010 database, experimental results show that the performances of the different systems evaluated with the proposed feature vectors give the best performance of 96.88% of IR. The use of GFCC does not give an improvement comparing to MFCC coefficients. The use of the proposed features is not efficient at this stage since there is no amelioration in speaker recognition performance with NIST SRE 2010 database. For that, several experiments were realized in order to detect the effectiveness of the proposed features at different dimensions. We deduce then the efficiency of the proposed features at a lower order, and we demonstrate that the use of only 13-dimensional MNGFCC and MVNGFCC feature vectors is very significant. Seen that we want to develop a realistic application in which we should taking into account of the memory and time complexity, the representation of the speech utterances with more reduced feature vectors reduce more the memory and time complexity of the system.

We examine then if we can also keep good performances. We decide to eliminate the delta and double delta coefficients of MFCC feature vectors and reduce the GFCC feature vectors so that 13-dimensional feature vectors are used for these features.

The results achieved with the different Reduced MFCC, GFCC feature vectors and the proposed MNMFCC, MVNMFCC, MNGFCC, and MVNGFCC features with 13-dimensional feature vectors that we call RMFCC, RGFCC, RMNMFCC, RMVNMFCC, RMNGFCC, and RMVNGFCC with 24 s for training and 6 s for testing with NIST SRE 2010 database are given with Fig. 4.

From these results, we can deduce that the use of Reduced MFCC (RMFCC) features is not efficient since it decreases the system performance and attain only 93.75% of correct IR. It is clear that the use of reduced MNMFCC and MVNMFCC slightly increases the performance which attains, respectively, 95.31% and 96.88% of IR, but it is not efficient in ameliorating the system performance.

The use of 13-dimensional GFCC, MNGFCC MVNGFCC features increases the system performance. In fact, the use of RGFCC and RMNGFCC features allow attaining 98.44% of correct IR. The use of RMVNGFCC is very efficient and succeeds to increase the system

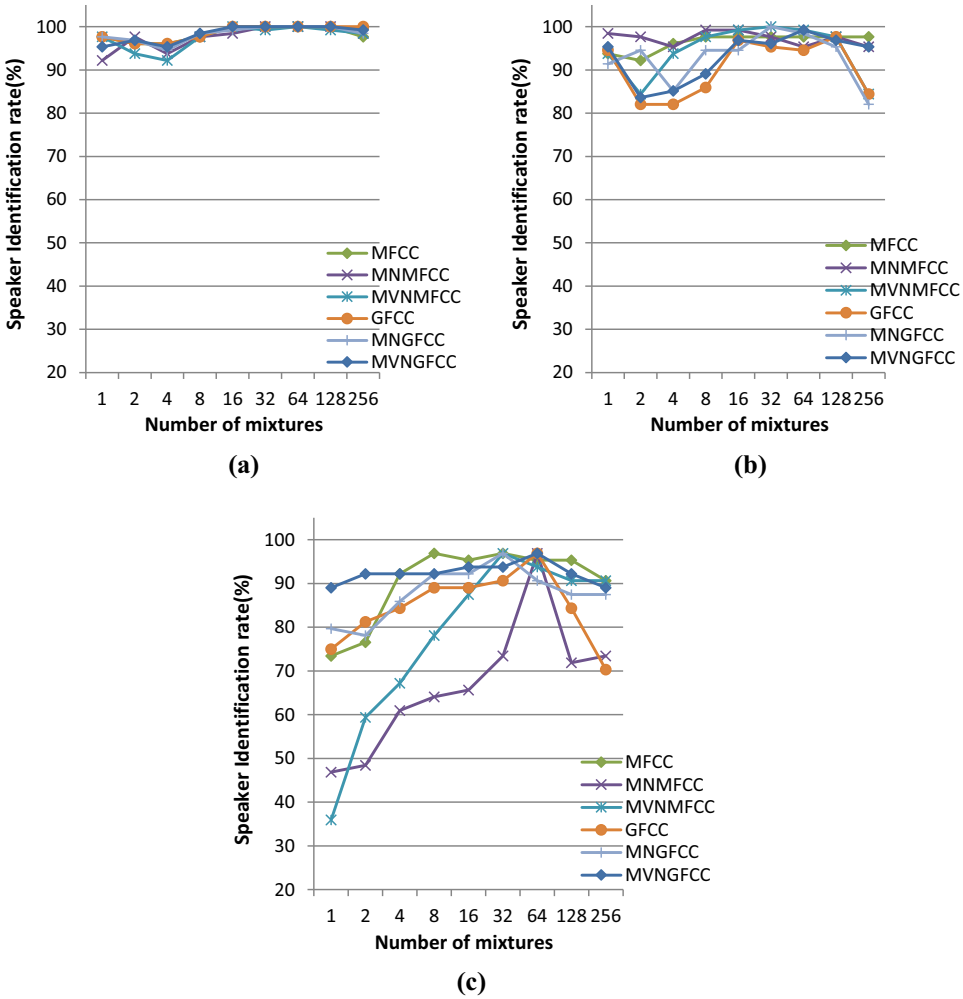

Fig. 3 Speaker IR with 24 s for training and 6 s for testing using the standard and the proposed features for the different databases **a** Speaker IR for TIMIT database, **b** speaker IR for NTIMIT database, **c** speaker IR for NIST SRE 2010 database

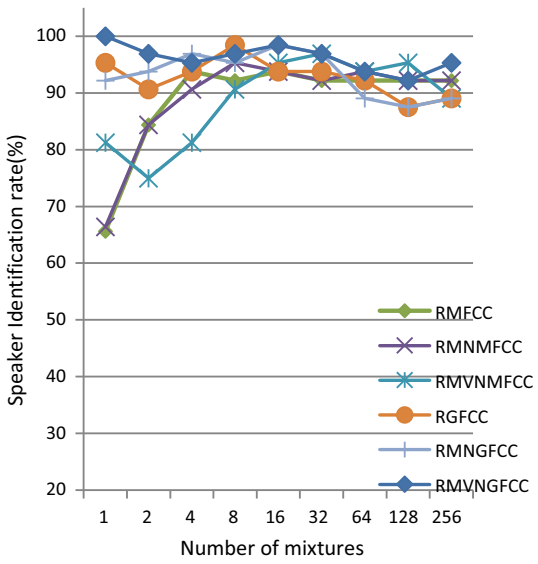

Fig. 4 Speaker IR with 24 s for training and 6 s for testing with the different reduced dimensional feature vectors with NIST SRE 2010 database

Table 1 The gain achieved with the different reduced dimensional feature vectors in term of correct Speaker IR for 24 s for training and 6 s for testing with NIST SRE 2010 database

Features	IR (%)	Gain (%)
MFCC	96.88	
GFCC	96.88	
RGFCC	98.44	1.56
RMNGFCC	98.44	1.56
RMVNGFCC	100	3.12

performance which reaches 100% of correct IR. In order to more highlight the efficiency of the proposed features, we give the achieved realized by the reduced proposed features as illustrated in Table 1.

These results clarify the potential superiority of the proposed RMVNGFCC features with i-vector-PLDA system. In fact, we realize a relatively important gain of about 1.5% with the proposed RGFCC and RMNGFCC features and 3% with the proposed RMVNGFCC features. The

effectiveness of the proposed system using the new RMNGFCC and RMVNGFCC feature vectors with the i-vector-PLDA classifier is further examined with more shortened training and testing data duration in the next section.

4.2.2 Speaker identification with short durations

In order to prove the efficiency of the proposed features, we evaluate the next set of experiments with more reduced training data. Hence, the following set of experiments is dealt with training speech utterances having duration of approximately 10 s and testing segments of 6 s for each speaker.

We begin by the evaluation of the performances of two baseline systems using, respectively, the standard 39 MFCC feature vectors and the standard GFCC feature vectors measured against the number of mixture components to highlight the effect of reducing the amount of training data on the systems' performance. Since the use of the reduced proposed feature vectors RGFCC, RMNGFCC, and RMVNGFCC allow achieving previously the best results, we evaluate then the performances of the proposed systems, using, respectively, the RGFCC, RMNGFCC, and RMVNGFCC feature vectors against the baseline systems with limited data duration. The results achieved from the different set of experiments for the different databases are given with Fig. 5.

The different results presented with the following Fig. 5 clearly indicate the effectiveness of the proposed systems compared to the baseline ones. With TIMIT database, we notice that the best MFCC-based baseline system performance is 95.31% with 16 mixtures. This result clearly explains the effect of reducing the amount of training data

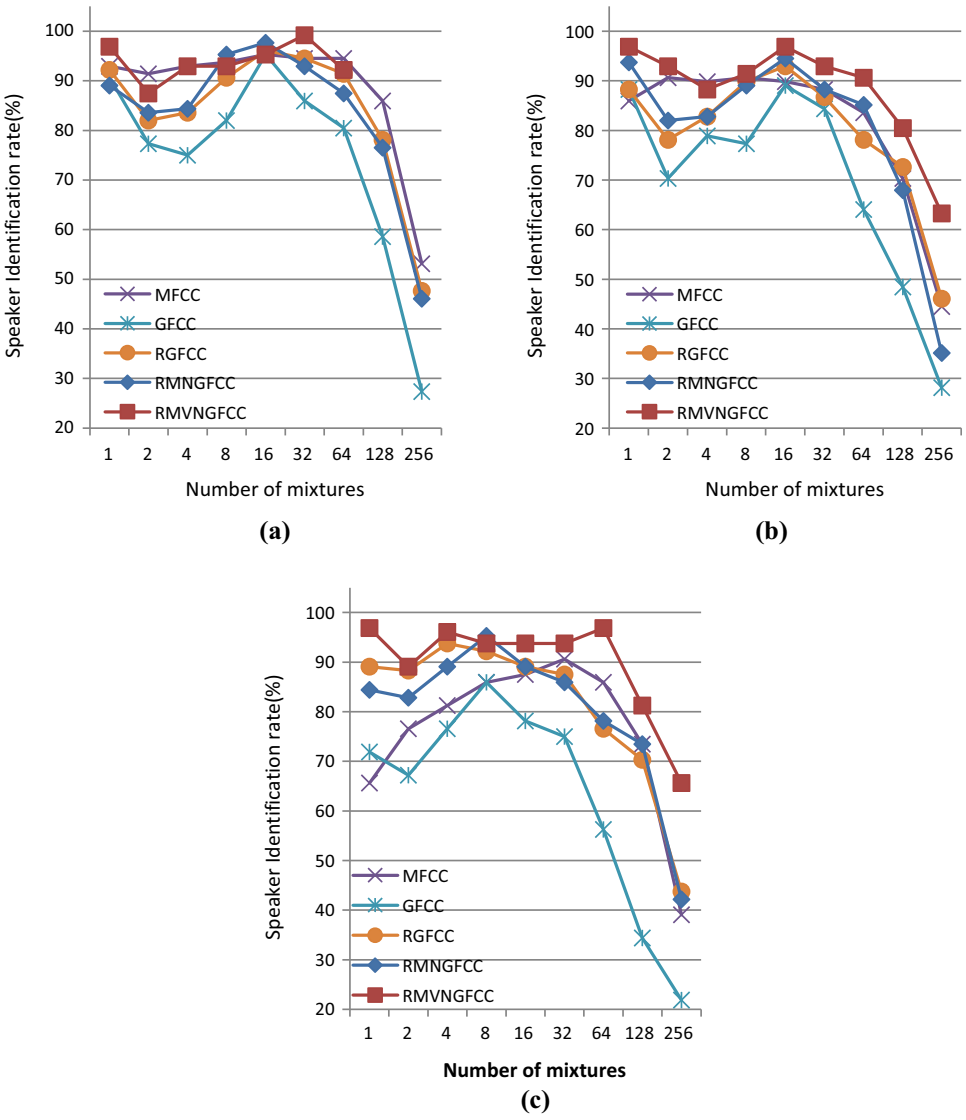

Fig. 5 Speaker IR with 10 s for training and 6 s for testing with baseline and proposed systems for the different databases. **a** Speaker IR with TIMIT database, **b** speaker IR with NTIMIT database, **c** speaker IR with NIST SRE 2010 database

on the system performance which achieved previously 100% with 24 s of training data. The use of the standard GFCC features maintains the same performance as with standard MFCC features in this case. The evaluation of the proposed speaker identification system with the RGFCC features brings further improvements to the system which could attain 96.09% of correct speaker IR. The system takes advantage from the use of RMNGFCC features and the best-achieved result is 97.66% with 16 mixtures. The superiority of the RMVNGFCC features is clearly observed since the system reached 99.22% with 32 mixtures.

For the speakers taken from the NTIMIT database, we can also observe from the experiments evaluated with baseline system using the standard MFCC feature vectors the effect of using reduced amount of training data on the system performance. In fact, the system attains only 90.63% of correct IR against 97.66% achieved previously with 24 s of training data. The use of the GFCC coefficients does not give an improvement in the system performance with the baseline system. The use of the RGFCC, RMNGFCC and RMVNGFCC feature vectors increased the system performance which achieved, respectively, 92.97%, 94.53%, and 96.88% of correct speaker IR.

For the experiments dealt with 2010 NIST SRE database, the speaker identification rates obtained with baseline system using MFCC feature vectors achieve 90.63% against 96.88% achieved previously with 24 s of training data which clarify the effect of reducing the amount of training data on the system performance. The use of the standard GFCC features provides no amelioration to the system performance.

The use of Reduced GFCC features increases the system performance which achieves an identification rate of 93.75% of speaker IR. The use of the new features with the proposed Speaker Recognition system succeed to increase the system performance which achieves 95.31% with RMNGFCC features and reached 96.88% with RMVNGFCC features in this case. We calculate the achieved gain imported by the proposed RMNGFCC features and the RMVNGFCC features against baseline GFCC features for the different databases, and we report them with Table 2.

Table 2 The gain achieved with the proposed RMNGFCC and RMVNGFCC features compared to baseline GFCC features in term of correct Speaker IR for 10 s for training and 6 s for testing with the different databases

Features	TIMIT Gain (%)	NTIMIT Gain (%)	NIST SRE 2010 Gain (%)
RGFCC	0.78	3.91	7.81
RMNGFCC	2.35	5.47	9.37
RMVNGFCC	3.91	7.82	10.94

From the results obtained above, we can conclude the superiority of both proposed RMNGFCC and RMVNGFCC features with the i-vector-PLDA system. In fact, we can obtain important gains that go beyond 3%, 7% and 10% with the proposed RMVNGFCC features, respectively, for TIMIT, NTIMIT, and NIST SRE 2010 databases.

Therefore, we can deduce that the use of the proposed features is very efficient, and a special attention can be made for the RMVNGFCC-based system and the RMNGFCC-based system which outperforms the use of the RGFCC-based system when a reduced amount of speech data is used for speaker identification.

The following set of experiments is dealt with more shortened speech data. In fact, we prepared a set of utterances having a length of 10 s, 8 s, 6 s, and even 4 s per speaker for the training task and utterances having a length of 3 s, 2 s, 1 s, and 0.5 s per speaker for the test task. The two baseline systems were then conducted with the standards MFCC and GFCC features evaluated with the i-vector-PLDA algorithm. Since the use of the RMNGFCC and RMVNGFCC features almost outperforms the use of the RGFCC features for the different databases, we decide to keep the RMNGFCC and RMVNGFCC features for the next set of experiments. Indeed, we evaluate our experiments with a first baseline system using the standards MFCC features and a second baseline system using the standards GFCC features and proposed systems using, respectively, the RMNGFCC features, the RMVNGFCC features and both concatenated RMNGFCC and RMVNGFCC features that we called FRMVGFCC features. Results achieved from the different set of experiences evaluated on the different databases are given with Figs. 6, 7, 8 and 9.

Overall, experimental results show that as the utterance length diminishes, the performance degrades with a decreasing identification rate. The results presented above clearly indicate how sensitive the systems are to the amount of training and testing data.

We can also remark that the GFCC-based approach often outperforms the MFCC-based approach for the different training and testing durations. In fact, with TIMIT database, with 8-s utterance training duration and 3-s data testing duration, the GFCC-based system performance achieved 96.88% against 95.31% obtained with the MFCC-based system. With more reduced data like 4 s utterance training duration, the GFCC-based system also outperforms the MFCC-based approach and achieved 92.19%, 88.28%, 65.63%, and 34.38% with, respectively, 3 s, 2 s, 1 s, and 0.5 s testing data duration against only 87.5%, 78.13%, 53.13%, and 32.81% obtained with the MFCC-based approach. The same remark is also validated with NTIMIT database for which the GFCC-based system

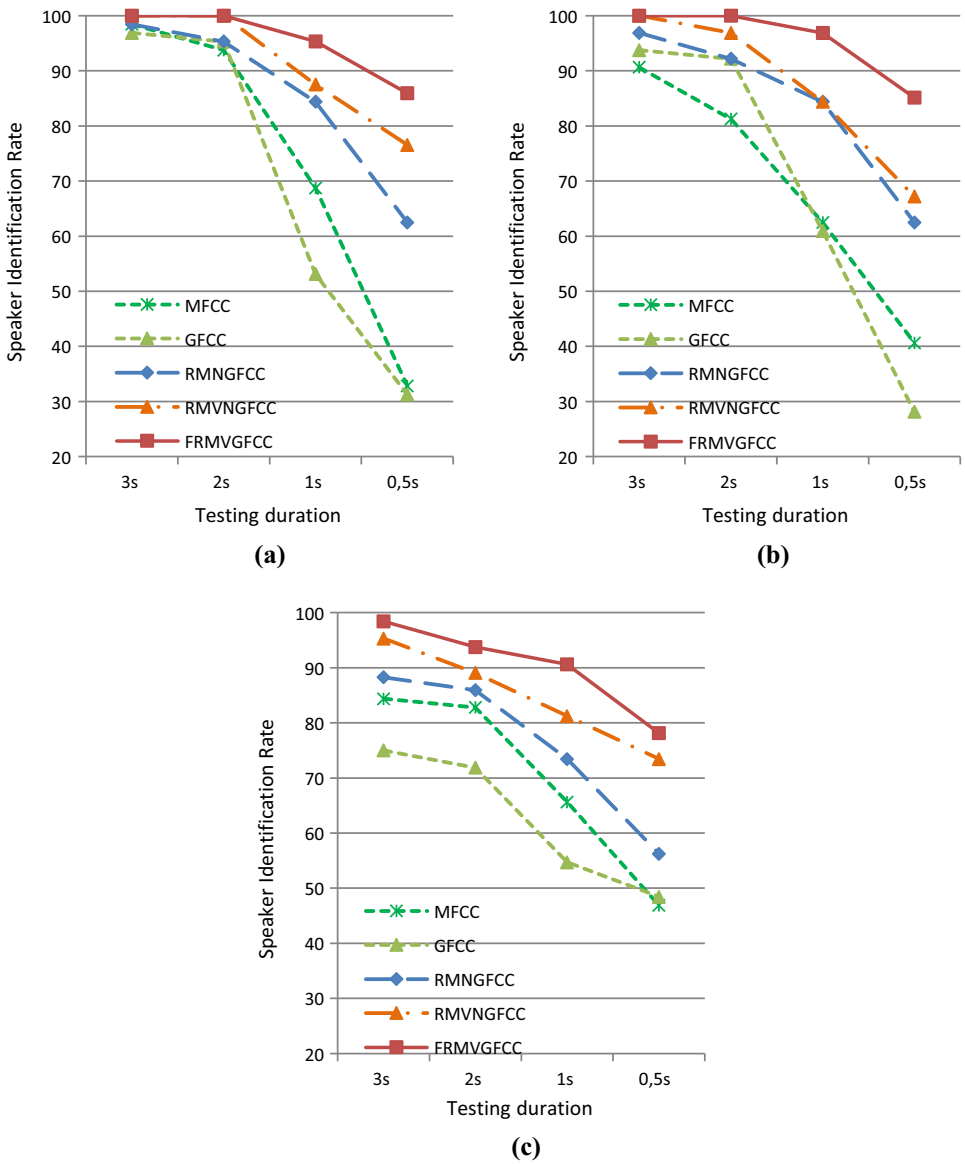

Fig. 6 Speaker IR for *10 s of training* and different testing durations with TIMIT, NTIMIT, and NIST SRE 2010 databases. **a** IR for TIMIT database, **b** IR for NTIMIT database, **c** IR for NIST SRE 2010 database

outperforms the MFCC-based system and achieved for example with 10 s utterance training duration 93.75% and 92.19% with, respectively, 3 s and 2 s testing data duration against only 90.63% and 81.25% with the MFCC-based system with more reduced training data like 4 s utterance training duration, the GFCC-based system gives 85.94%, 78.91% 53.91%, and 28.13% with, respectively, 3 s, 2 s, 1 s, and 0.5 s testing data duration against only 76.56%, 68.75%, 39.06%, and 18.75% obtained with the MFCC-based approach. With NIST SRE 2010 database, we can also remark that the use of the GFCC features can favorite the system performance and gives for example 57.81% with 4 s of training and 3 s of testing data duration against only 54.69% obtained with the MFCC-based approach.

If we compare between the proposed systems using the RMNGFCC and RMVNGFCC features and baseline systems, we can clearly remark the effectiveness of these features with the increase of the obtained IR for the different databases. In fact, with TIMIT database, the proposed system using the RMNGFCC features achieves for example 98.44% with 10 s of training duration and 3 s s testing data duration. The use of RMVNGFCC features further increased the performances which attain 100% of correct IR with 10 s of training duration, 3 s and even 2 s of testing data duration.

The use of the proposed features is also very beneficial for more reduced training and testing data duration. In fact, the use of the RMNGFCC achieved for 4 s of training 90.63%, 92.97%, 82.81%, and 59.38% with, respectively, 3 s, 2 s, 1 s and 0.5 s testing data duration against only 87.5%, 78.13%, 53.13%, and 32.81% with the MFCC-based approach. The use of RMVNGFCC features clearly

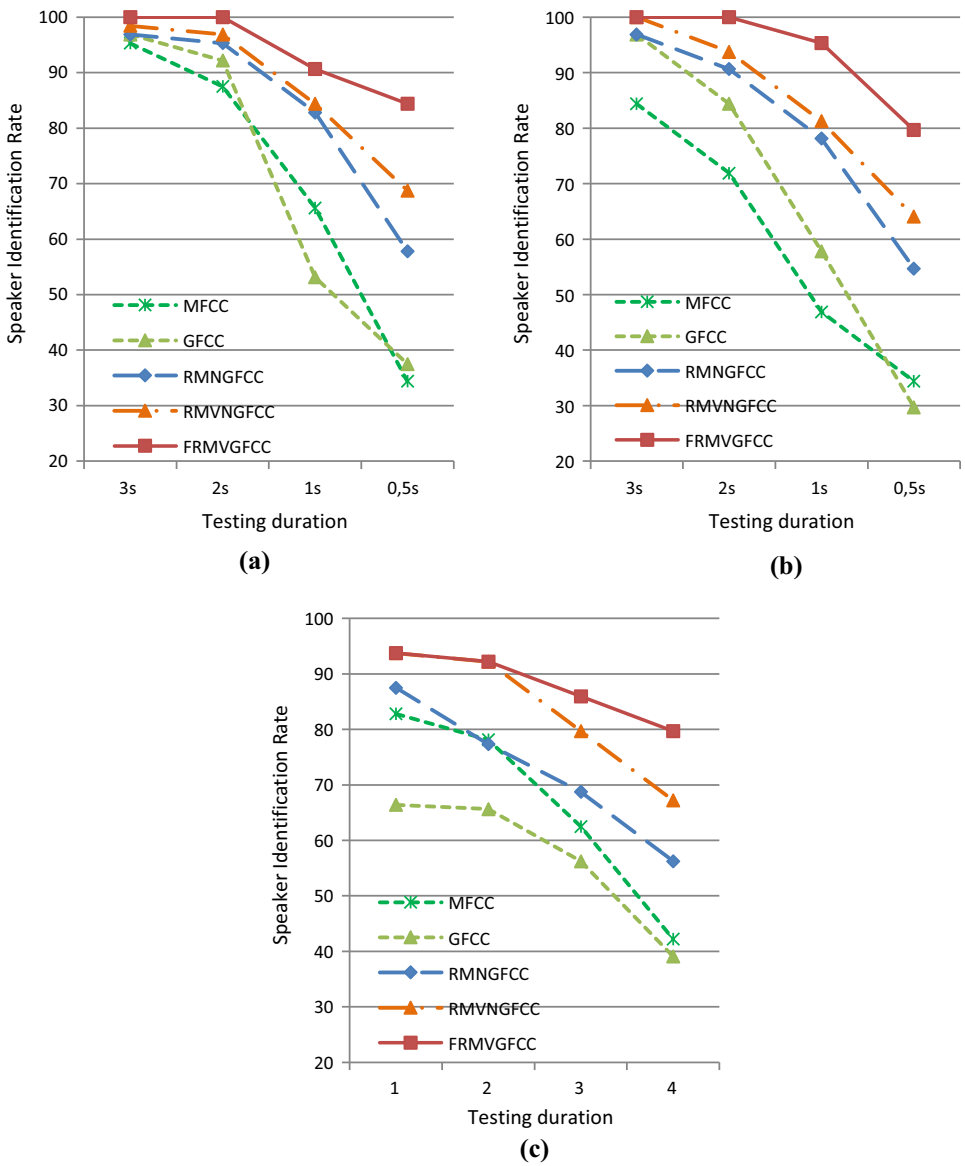

Fig. 7 Speaker IR for *8 s of training* and different testing durations with TIMIT, NTIMIT, and NIST SRE 2010 databases. **a** IR for TIMIT database, **b** IR for NTIMIT database, **c** IR for NIST SRE 2010 database

proves the outperformance of the proposed system which achieved, respectively, 98.44%, 92.19%, 81.25% and 64.06% with, respectively, 3 s, 2 s, 1 s, and 0.5 s of testing data duration.

With NTIMIT database, we notice a great improvement for the proposed system performance. For example, the proposed system succeeds to achieve 100% of speaker identification with 10 s of training and 3 s of testing with RMVNGFCC features and 96.88% with RMNGFCC features against only 90.63% and 93.75% of speaker IR with standard MFCC and GFCC features.

Similar observation can also be made for more shortened training and testing data. In fact, with 4 s of training and 3 s test utterance durations, the performances are proved and the proposed features increase the speaker IR from 76.56% and 85.94% with the standard MFCC and GFCC features to 89.06% and 92.19% with the proposed RMNGFCC and RMVNGFCC features.

The results obtained with NIST SRE 2010 database clearly prove the superiority of the proposed system. In applying RMVNGFCC features with the proposed system, the system gives additional improvement compared to the results obtained with the proposed system using RMNGFCC features. In fact, for example, with 10 s of training and 0.5 s of testing, the best-achieved results with the proposed system were 56.25% and 73.44%, respectively, with RMNGFCC and RMVNGFCC features against only 46.88% and 48.44% with the standard MFCC and GFCC features. For 10 s of speech training duration and 3 s test utterance durations, the proposed system improve the system performance which increases from 84.38% and

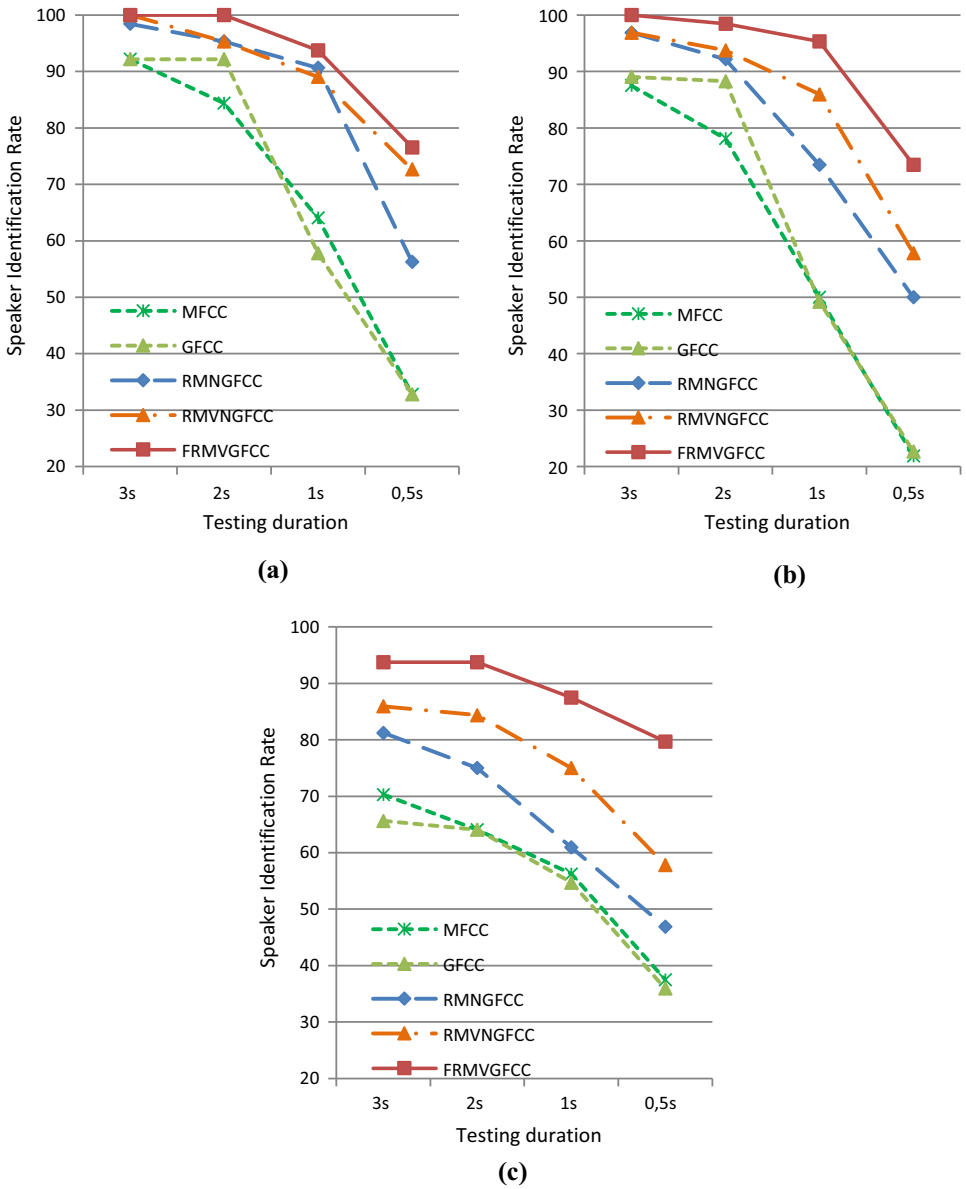

Fig. 8 Speaker IR for *6 s of training* and different testing durations with TIMIT, NTIMIT, and NIST SRE 2010 databases. **a** IR for TIMIT database, **b** IR for NTIMIT database, **c** IR for NIST SRE 2010 database

75% with the standard features and reaches 88.28% and 95.31% with the proposed features.

For more reduced training duration like 8 s, the performance decrease and the identification rate fall to 42.19% with the reduction of the testing duration to 0.5 s with the standard MFCC features. Hence, we can also observe better performance with the proposed features which enhance the results and give 56.25% and 67.19%, respectively, with RMNGFCC and RMVNGFCC features. We can remark inferior performance obtained with 6 s of training. In fact, with baseline systems we have for example only 70.31% for 3 s of testing with the standard MFCC features. The proposed features succeed to increase the system performance which reaches 81.25% and 85.94%, respectively, with RMNGFCC and RMVNGFCC features.

The use of 4 s only of training data diminishes further the system performance and gives 54.69% and 57.81% of identification with 3 s of testing with the standard MFCC and GFCC features. The proposed features ameliorate the system performance which reaches 75.78% and 79.69%, respectively, with RMNGFCC and RMVNGFCC features.

These results prove how sensible the speaker identification system is to the duration of training and testing speech segments. For that, the use of the proposed features is very essential to compensate the decreasing performance caused by the use of short utterances especially in non-controlled conditions.

Motivated by their superior performance with regard to the other features as demonstrated with experimental results given above, we decide to take further advantage of

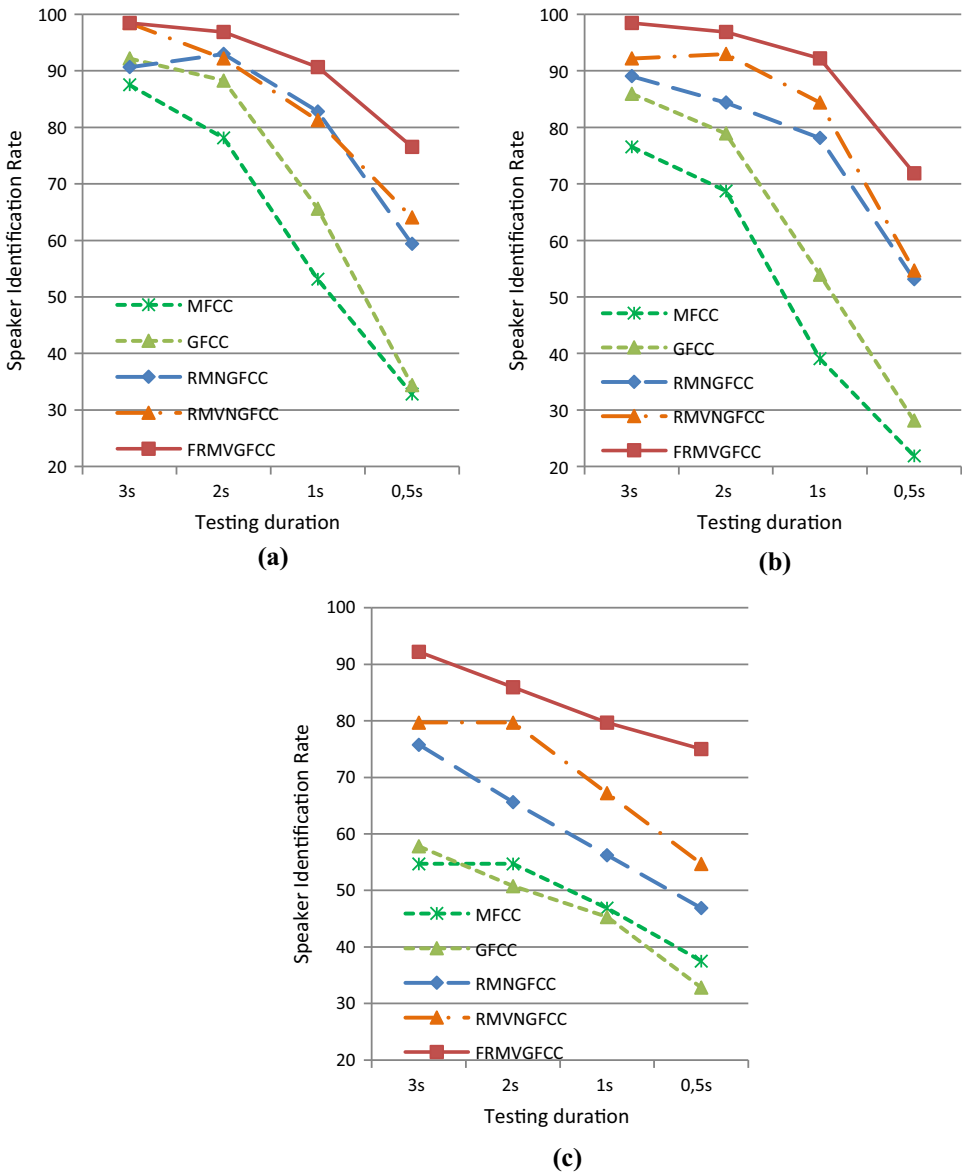

Fig. 9 Speaker IR for 4 s of training and different testing durations with TIMIT, NTIMIT and NIST SRE 2010 databases. **a** IR for TIMIT database, **b** IR for NTIMIT database, **c** IR for NIST SRE 2010 database

the potential superiority of RMNGFCC and RMVNGFCC features when limited amounts of training and testing data are available. Hence, we decide to use both of the two proposed features. The performance of the new proposed speaker identification system which depends on both combined RMNGFCC and RMVNGFCC features called FRMVGFCC features was evaluated following the protocol described previously with the new resulting feature vector so we can clearly demonstrate our contribution to short utterance speaker identification. In fact, experimental results given above clearly highlight that the use of the proposed approach helps to favor the system performance for the different databases. In fact, if we focus for example on NIST SRE 2010 database, we found that the best result achieved with the proposed resulting system is 98.44% with 10 s of training and 3 s of testing when both features are used against 95.31% with RMVNGFCC features only and 88.28% with RMNGFCC features. With 0.5 s per speaker for the test task, the best result achieved with the proposed system with 10 s of training was 78.13% of identification. However, it was only 73.44% of identification with the RMVNGFCC-based system.

We achieve also 85.94% of identification with 8 s of training and 1 s of testing with the proposed system. However the best performance was only 79.69% with the RMVNGFCC-based system.

The use of the proposed system with more reduced training data like 6 s help to improve the system performance which gives 93.75% for 3 s of testing data and 79.69% for 0.5 s of testing which outperform the use of the outperformed RMVNGFCC features only which gives, respectively, 85.94 and 57.81% in these cases.

The use of 4 s of training data further diminishes the proposed system performance and gives 92.18% with 4 s for training and 3 s of testing and 75% with 4 s of training and 0.5 s of testing. Despite this reduction, this system still outperform the RMVNGFCC-based system which achieves only 79.69% of identification with 4 s of training and 3 s of testing and only 54.69% of identification with 0.5 s of testing.

The same remarks are also clear for the other databases. In fact, with TIMIT database, the proposed system further increase the speaker identification performance which passes for example from 84.38% and 87.5% respectively with RMNGFCC and RMVNGFCC features to 95.31% with the proposed system for 8 s of training and 1 s of testing.

The proposed system also increase the recognition performance for more reduced speech data achieve 76.56% with only 4 s of training and 0.5 s of testing against only 59.38% and 64.06% respectively with RMNGFCC and RMVNGFCC features.

The superior outperformance is also remarkable for NTIMIT database. In fact, for 4 s of training and 0.5 s of testing, the proposed system achieves 71.88% of correct IR against only 53.13% and 54.69% respectively with RMNGFCC and RMVNGFCC features.

The effectiveness of the proposed features and the proposed system is highly efficient for the different databases. In fact, for example, if we focus on reduced data like 4 s of training and 0.5 s of testing, we found that the proposed system increase the IR from 32.81% and 34.38% with standard MFCC and GFCC features to 76.56% with the proposed system with TIMIT database.

The same remark is also valid for NTIMIT Database since the proposed system achieves 71.88% against only 21.88% and 28.13% respectively with standard MFCC and GFCC features. For NIST SRE 2010 database, the best-achieved result with the proposed resulting system for 4 s of training and 0.5 s of testing is 75% against only 37.5% and 32.81% with standard MFCC and GFCC features. For 4 s of training and 3 s of testing, the proposed system achieved 92.19% against only 54.69% and 57.81% with standard MFCC and GFCC features.

In order to further highlight the contribution imported by the proposed new features, we calculate the achieved gain realized by the proposed RMNGFCC, RMVNGFCC, and FRMVGFCC features compared to baseline GFCC features for the limited 4 s of training utterance duration and the different short testing durations for the different databases, and we report them with Tables 3, 4 and 5.

From the results given above, we can conclude the effectiveness of the proposed RMNGFCC, RMVNGFCC, and FRMVGFCC features with the i-vector-PLDA system. In fact, the superiority of the new features is remarkable with the different databases. For TIMIT database, important gains that surpass 6% are observed for 3 s of testing data duration. With the use of 2 s of testing utterance duration, the proposed FRMVGFCC features achieved a gain that go beyond 8% of correct identification rate. Using shorter testing utterances like 1 s of testing segments' duration demonstrates the usefulness of the proposed features which achieved a gain of about 17% and 15%, respectively, with the proposed RMNGFCC and RMVNGFCC features. The proposed FRMVGFCC features achieved an important gain of 25% in this case. The same remark is also validated with 0.5 s of testing utterance duration where the proposed FRMVGFCC features achieved an important gain of about 44% of correct identification rate.

We can also appreciate the outperformance of the proposed features for the NTIMIT database. In fact, for 3 s of testing data duration, the proposed FRMVGFCC features

Table 3 The gain achieved with the proposed RMNGFCC, RMVNGFCC and FRMVGFCC features compared to baseline GFCC features in term of correct Speaker IR for 4 s of training and different testing durations with TIMIT database

Features	Gain (%)			
	3 s	2 s	1 s	0.5 s
RMNGFCC	–	4.69	17.18	25
RMVNGFCC	6.25	3.91	15.62	29.68
FRMVGFCC	6.25	8.6	25	42.18

Table 4 The gain achieved with the proposed RMNGFCC, RMVNGFCC and FRMVGFCC features compared to baseline GFCC features in term of correct Speaker IR for 4 s of training and different testing durations with NTIMIT database

Features	Gain (%)			
	3 s	2 s	1 s	0.5 s
RMNGFCC	3.12	5.47	24.22	25
RMVNGFCC	6.25	14.06	30.47	26.56
FRMVGFCC	12.5	17.97	38.28	43.75

Table 5 The gain achieved with the proposed RMNGFCC, RMVNGFCC and FRMVGFCC features compared to baseline GFCC features in term of correct Speaker IR for 4 s of training and different testing durations with NIST SRE 2010 databases

Features	Gain (%)			
	3 s	2 s	1 s	0.5 s
RMNGFCC	17.97	14.85	10.94	14.07
RMVNGFCC	21.88	28.91	21.88	21.88
FRMVGFCC	34.38	35.16	34.38	42.19

achieved an important gain that surpasses 12% of correct identification rate compared to the GFCC baseline features. For the use of more shortened testing data like 2 s of testing utterance duration, the proposed RMNGFCC achieved a gain of more than 5% of correct identification rate, the RMVNGFCC features achieved an important gain of about 14% of correct identification rate and the proposed FRMVGFCC features achieved higher gain of about 18% of correct identification rate compared to the baseline features. These features are also very efficient with more shortened testing utterance duration. In fact, the proposed FRMVGFCC features improve the system performance with more than 38% of correct identification rate compared to baseline features with 1 s of testing utterance duration and more than 43% of correct identification rate compared to baseline features with 0.5 s of testing data duration.

The evaluations with the NIST SRE 2010 database permit also to prove the efficiency of the new features. In fact, with 3 s of testing data duration the achieved gains outperform 17%, 21%, and 34%, respectively, with the proposed RMNGFCC, RMVNGFCC, and FRMVGFCC features compared to the standards GFCC features. These features prove also their effectiveness with more reduced speech utterance durations and the achieved gain go beyond 35% of correct identification rate with the use of 2 s of testing utterance duration with the FRMVGFCC features. The same remark is also appreciated with 1 s of testing data duration since the RMNGFCC, RMVNGFCC, and FRMVGFCC features succeed to increase the system performance with about 11%, 22%, and 35% compared to the baseline system. The contribution of the proposed system using the FRMVGFCC features is thereby validated with 0.5 s of testing utterance duration and the proposed system increase the system performance with an important gain of 42.19% of correct identification rate in this case.

In this way, we can conclude that the proposed system using FRMVGFCC features is very efficient since it can achieves higher identification rate with regard to baseline approaches for both reduced training and testing utterance with clean and noisy speech utterances. In fact, the given evaluations showed that our approach achieved remarkable results with short test utterances and limited data in the training phase. Hence, the proposed features succeed to store the maximum of information about the speaker's characteristics. Seen that we use reduced feature vectors compared to baseline and standard approaches, the proposed system presents an efficient solution to overcome and mollify also the constraints related to the memory and computational resource limitation in realistic applications, and hence makes possible the use of large datasets containing many speakers without the need of incorporating additional, lengthy and complicated algorithms requiring more time and memory space.

5 Conclusions and perspectives

Although many recent advances and successes have been achieved with speech researchers, the challenges of providing effective robust speaker identification on short utterances remain a key consideration when deploying automatic speaker recognition, as many real-world applications often have access to only limited duration speech data recorded under uncontrolled conditions.

This paper has introduced and evaluated the use of an enhanced i-vector-PLDA system based on new reduced dimensional feature vectors for robust text-independent speaker identification. The proposed system was specifically evaluated for speaker identification purpose using short-duration utterances for both training and testing task obtained from unconstrained speech transmitted over noise encountered telephone channels. This proposed method has focused on the formulation of a new approach looking for new information able to facilitate the identification of speakers with much reduced speech information. We prove that this method is suitable for a realistic speaker recognition application. In fact we do not need to use a large amount of training dataset as in traditional algorithms. Besides, we don't require long test utterances to recognize the speaker. Moreover, there is no need to incorporate lengthy and complicated calculations to handle the situations of having small amounts of speech data. This is an interesting advantage especially for realistic applications that need to reduce the computational and time complexity of the system and so the memory size of the system. Future work will also consider the performance of the proposed system with other features or applications.

Compliance with ethical standard

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

References

1. Liu JC, Leu FY, Lin GL, Susanto H (2018) An MFCC-based text-independent speaker identification system for access control. Concur Comput Pract Exp 30(2):e4255
2. Togneri R, Pullella D (2011) An overview of speaker identification: accuracy and robustness issues. IEEE Circuits Syst Mag 11(2):23–61
3. Dişken G, Tüfekçi Z, Saribulut L, Çevik U (2017) A review on feature extraction for speaker recognition under degraded conditions. IETE Tech Rev 34(3):321–332
4. Larcher A, Bonastre JF, Mason JS (2008) Short utterance-based video aided speaker recognition. In: 2008 IEEE 10th workshop on multimedia signal processing, pp 897–901. IEEE

5. Chang J, Wang D (2017) Robust speaker recognition based on DNN/i-vectors and speech separation. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5415–5419. IEEE
6. Ranjan S, Misra A, Hansen JH (2017) Curriculum learning based probabilistic linear discriminant analysis for noise robust speaker recognition. Proc Interspeech 2017:3717–3721
7. Krishnamoorthy P, Jayanna HS, Prasanna SM (2011) Speaker recognition under limited data condition by noise addition. Expert Syst Appl 38(10):13487–13490
8. 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
9. Chakroun R, Frikha M, Zouari LB (2018) New approach for short utterance speaker identification. IET Signal Process 12(7):873–880
10. Fatima N, Zheng TF (2012) Short utterance speaker recognition a research agenda. In: International conference on systems and informatics (ICSAI)
11. Liu Z, Wu Z, Li T, Li J, Shen C (2018) GMM and CNN hybrid method for short utterance speaker recognition. IEEE Trans Ind Inf 14(7):3244–3252
12. Park SJ, Yeung G, Kreiman J, Keating PA, Alwan A (2017) Using voice quality features to improve short-utterance, text-independent speaker verification systems. Proc Interspeech 2017:1522–1526
13. Khosravani A, Homayounpour MM (2018) Nonparametrically trained PLDA for short duration i-vector speaker verification. Comput Speech Lang 52:105–122
14. Matza A, Bistritz Y (2014) Skew Gaussian mixture models for speaker recognition. IET Signal Process 8(8):860–867
15. Motlicek P, Dey S, Madikeri S, Burget L (2015) Employment of subspace gaussian mixture models in speaker recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4445–4449
16. 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
17. Saeidi R, Alku P (2015) Accounting for uncertainty of i-vectors in speaker recognition using uncertainty propagation and modified imputation. In: Proceedings of Interspeech, vol 2015
18. Sholokhov A, Sahidullah M, Kinnunen T (2018) Semi-supervised speech activity detection with an application to automatic speaker verification. Comput Speech Lang 47:132–156
19. 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 (TASLP) 24(6):1129–1139
20. Reynolds D, Quatieri T, Dunn R (2000) Speaker verification using adapted Gaussian mixture models. Digit Signal Process 10(1–3):19–41
21. Li S, Karatzoglou A, Gentile C (2016) Collaborative filtering bandits. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 539–548
22. Korda N, Szörényi B, Shuai L (2016) Distributed clustering of linear bandits in peer to peer networks. In: Journal of machine learning research workshop and conference proceedings, vol 48. International Machine Learning Societ, pp 1301–1309
23. Li S (2016) The art of clustering bandits. Doctoral dissertation, Università degli Studi dell'Insubria
24. Dehak N, Kenny P, Dehak R, Dumouchel P, Ouellet P (2011) Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang Process 19(99):788–798
25. Sarkar A, Matrouf D, Bousquet P, Bonastre J (2012) Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. In: Thirteenth annual conference of the international speech communication association, INTERSPEECH, pp 2662–2665
26. Kanagasundaram A, Vogt R, Dean D, Sridharan S, Mason M (2011) I-vector based speaker recognition on short utterances. In: Proceedings of Interspeech, Florence, Italy, 2011, pp 2341–2344
27. Mandasari MI, McLaren M, van Leeuwen DA (2011) Evaluation of i-vector speaker recognition systems for forensic application. In: Proceedings of Interspeech. ISCA, Firenze
28. Hasan T, Saeidi R, Hansen JHL, van Leeuwen DA (2013) Duration mismatch compensation for i-vector based speaker recognition systems. In: Proceedings of IEEE ICASSP, Vancouver, Canada
29. The NIST year 2012 speaker recognition evaluation plan (2012). [online] Available: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf
30. Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: from features to supervectors. Speech Commun 52(1):12–40
31. Zhang WQ, Zhao J, Zhang WL, Liu J (2014). Multi-scale kernels for short utterance speaker recognition. In: The 9th international symposium on Chinese spoken language processing. IEEE, pp 414–417
32. Fauve B, Evans N, Mason J (2008) Improving the performance of text-independent short duration SVM-and GMM-based speaker verification. In: Proceedings of Odyssey, Stellenbosch, South Africa
33. McLaren M, Vogt R, Baker B, Sridharan S (2010) Experiments in SVM-based speaker verification using short utterances. In: Proceedings of Odyssey workshop 2010
34. Lan Y, Hu Z, Soh YC, Huang GB (2013) An extreme learning machine approach for speaker recognition. Neural Comput Appl 22(3–4):417–425
35. Heigold G, Moreno I, Bengio S, Shazeer N (2016) End-to-end text-dependent speaker verification. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5115–5119
36. Zhang SX, Chen Z, Zhao Y, Li J, Gong Y (2017) End-to-end attention based text-dependent speaker verification. arXiv preprint arXiv:1701.00562
37. Variani E, Lei X, McDermott E, Moreno IL, Gonzalez-Dominguez J (2014) Deep neural networks for small footprint text-dependent speaker verification. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4052–4056
38. Heigold G, Moreno I, Bengio S, Shazeer N (2016) End-to-end-text-dependent speaker verification. In: 2016 IEEE international conference on Acoustics, speech and signal processing (ICASSP). IEEE, pp 5115–5119
39. Zhang C, Koishida K (2017) End-to-end text-independent speaker verification with triplet loss on short utterances. In: Interspeech, Copyright © 2017 ISCA, August 20–24, Stockholm, Sweden, pp 1487–1491. https://doi.org/10.21437/Interspeech.2017-1608
40. Snyder D, Ghahremani P, Povey D, Garcia-Romero D, Carmiel Y, Khudanpur S (2016) Deep neural network-based speaker embeddings for end-to-end speaker verification. In: 2016 IEEE spoken language technology workshop (SLT), IEEE, pp 165–170
41. Bhattacharya G, Alam MJ, Kenny P (2017) Deep speaker embeddings for short-duration speaker verification. In: Interspeech, Copyright© 2017 ISCA, August 20–24, Stockholm, Sweden, pp 1517–1521. https://doi.org/10.21437/Interspeech.2017-1575
42. Kanagasundaram A, Vogt R, Dean D, Sridharan S (2012) PLDA based speaker recognition on short utterances. In: The speaker and language recognition workshop (Odyssey 2012), ISCA, 2012

43. Kanagasundaram A, Dean D, Sridharan S (2014) Improving PLDA speaker verification with limited development data. In: IEEE international conference on acoustics, speech and signal processing
44. Rahman MH, Kanagasundaram A, Himawan I, Dean D, Sridharan S (2018) Improving PLDA speaker verification performance using domain mismatch compensation techniques. Comput Speech Lang 47:240–258
45. Cumani S, Plchot O, Laface P (2014) On the use of i-vector posterior distributions in probabilistic linear discriminant analysis. IEEE Trans Audio Speech Lang Process 22(4):846–857
46. Ganapathy S, Mallidi SH, Hermansky H (2014) Robust feature extraction using modulation filtering of autoregressive models. IEEE Trans Audio Speech Lang Process 22(8):1285–1295
47. Zhao X, Wang Y, Wang D (2014) Robust speaker identificat ion in noisy and reverberant conditions. IEEE Trans Audio Speech Lang Process 22(4):836–845
48. Yu C, Liu G, Hahm S, Hansen JHL (2014) Uncertainty propagation in front end factor analysis for noise robust speaker recognition. In: Proceedings of the 39th ICASSP, Florence, Italy, pp 4017–4021
49. Hurmalainen A, Saeidi R, Virtanen T (2015) Noise robust speaker recognition with convolutive sparse coding. In: Sixteenth annual conference of the international speech communication association
50. Lei Y, McLaren M, Ferrer L, Scheffer N (2014) Simplified VTS-based i-vector extraction in noise-robust speaker recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 4037–4041. IEEE
51. Kheder WB, Matrouf D, Bousquet PM, Bonastre JF, Ajili M (2017) Fast i-vector denoising using MAP estimation and a noise distributions database for robust speaker recognition. Comput Speech Lang 45:104–122
52. Ming J, Hazen TJ, Glass JR, Reynolds DA (2007) Robust speaker recognition in noisy conditions. IEEE Trans Audio Speech Lang Process 15(5):1711–1723
53. Lei Y, Burget L, Scheffer N (2013)A noise robust i-vector extractor using vector Taylor series for speaker recognition. In: Proceedings of the 38th ICASSP, Vancouver, BC, Canada, 2013, pp 6788–6791
54. Alku P, Saeidi R (2017) The linear predictive modeling of speech from higher-lag autocorrelation coefficients applied to noise-robust speaker recognition. IEEE/ACM Trans Audio Speech Lang Process 25:1606–1617
55. Liu X, Sadeghian R, Zahorian SA (2017) A modulation feature set for robust automatic speech recognition in additive noise and reverberation. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5230–5234
56. Zhao X, Shao Y, Wang DL (2012) CASA based robust speaker identification. IEEE Trans Audio Speech Lang Process 20(51):608–1616
57. Venkatesan R, Ganesh AB (2018) Binaural classification-based speech segregation and robust speaker recognition system. Circuits Syst Signal Process 37(8):3383–3411
58. Fedila, M, Bengherabi M, Amrouche A (2018) Gammatone filterbank and symbiotic combination of amplitude and phase-based spectra for robust speaker verification under noisy conditions and compression artifacts. Multimedia Tools Appl 77(13):16721–16739
59. Atal B (1974) Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J Acoustic Soc Am 55:1304
60. Mammone R, Zhang X, Ramachandran R (1996) Robust speaker recognition: a feature-based approach. IEEE Signal Process Mag 13(5):58–71
61. Reynolds D (1994) Experimental evaluation of features for robust speaker identification. IEEE Trans Speech Audio Process 2(4):639–643
62. Sheikhan M, Gharavian D, Ashoftedel F (2012) Using DTW neural–based MFCC warping to improve emotional speech recognition. Neural Comput Appl 21(7):1765–1773
63. Turner C, Joseph A (2015) A wavelet packet and mel-frequency cepstral coefficients-based feature extraction method for speaker identification. Procedia Comput Sci 61:416–421
64. Shahamiri SR, Salim SSB (2014) Artificial neural networks as speech recognisers for dysarthric speech: identifying the best-performing set of MFCC parameters and studying a speaker-independent approach. Adv Eng Inf 28(1):102–110
65. Ali H, Tran SN, Benetos E, Garcez ASDA (2018) Speaker recognition with hybrid features from a deep belief network. Neural Comput Appl 29(6):13–19
66. Young S, Kershaw D, Odell J, Ollason D, Valtchev V, Woodland P (2002) Hidden Markov model toolkit (HTK) version 3.4 user's guide
67. 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 7:27874–27882
68. Islam MA, Jassim WA, Cheok NS, Zilany MSA (2016) A robust speaker identification system using the responses from a model of the auditory periphery. PLoS ONE 11(7):e0158520
69. Zhao X, Shao Y, Wang D (2012) CASA-based robust speaker identification. Audio Speech Lang Process IEEE Trans 20(5):1608–1616
70. Zhao X, Wang D (2013) Analyzing noise robustness of MFCC and GFCC features in speaker identification. In: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 7204–7208
71. Shao Y, Wang D (2008) Robust speaker identification using auditory features and computational auditory scene analysis. In: IEEE international conference on acoustics, speech and signal processing, 2008. ICASSP 2008. IEEE, pp 1589–1592
72. Kenny P (2010) Bayesian speaker verification with heavy-tailed priors. In: Proceedings of odyssey speaker and language recognition workshop
73. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL (1993) DARPA TIMIT acoustic phonetic continuous speech corpus CDROM. NIST
74. Feng L, Hansen LK (2005) A new database for speaker recognition. Informatics and mathematical modeling. Technical University of Denmark, DTU
75. Reynolds DA (1995) Automatic speaker recognition using gaussian mixture speaker models. Linc Lab J 8(2):173–192
76. Jankowski C, Kalyanswamy A, Basson S, Spitz J (1990) NTI-MIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. ICASSP
77. The NIST Year 2010 Speaker Recognition Evaluation Plan (2010). http://www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplan-r6.pdf