International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 12, No.1 (Aug-2022)

https://dx.doi.org/10.12785/ijcds/120143

A Hybrid of Deep Neural Network and eXtreme Gradient Boosting for Automatic Speaker Identification

Dehia Abdiche[1] **and Khaled Harrar**[1]

[1] *LIST laboratory, University M'Hamed Bougara of Boumerdes, Algeria*

Received 5 Nov.2021, Revised 15 Mar. 2022, Accepted 15 Jul. 2022, Published 6 Aug. 2022

Abstract: This work consists of deploying a system for Speaker Identification (SI). SI is a system of recognition of the speaker's speech signal. The most important thing in SI is to have a system that is able to extract and learn discriminative and relevant features for classification. Most research on SI has shown the effectiveness of Perceptual Linear Predictive (PLP) and Mel-Frequency Cepstral Coefficients (MFCC). Nevertheless, these extraction techniques exhibit identification errors when the speech signal is complex. To overcome this problem, this study proposes two features extraction techniques. The first technique uses Mel-Frequency Energy Coefficients (MFEC), the second technique is a hybrid approach combining MFEC and Convolutional Neural Network (CNN) used as features extractors. SI was performed using the features derived from the speech signals in the Voxforge database by both classifiers, namely CNNs and eXtreme Gradient Boosting (XGBoost). The proposed hybrid model using XGBoost-CNN achieved an accuracy of 99.45% demonstrating the effectiveness of this combination for SI. Moreover, a comparative study was carried out and revealed that the proposed model provided promising results and outperformed the existing methods in the literature using the Voxforge database.

Keywords: Speaker identification, MFEC, CNN, XGBoost.

1. Introduction

Over time, technology is becoming much more present in our life, and human beings have become accustomed to manipulating these tools with great ease, they will even become dependent on them. Biometrics is one of these methods that are becoming indispensable to man. There are several forms of biometrics, including face,finger print, retina, voice, etc. Voice biometrics is one of the most important technologies, it has the advantage of being able to be used remotely using a simple phone for example. Voice biometrics is closely related to speaker recognition [1], [2], [3], [4], [5], [6], [7], which considers the unique features found in a speaker's speech signal. Speaker recognition is expanding in various applications, such as fast access control systems, [8], [9], [10], [11]. online transactions, multimedia and personalization, speech data management, computer access control [12], mobile banking, and mobile shopping [13]. Speaker recognition counts authentication, verification, and identification.

The most important step in the automatic speaker identification process is features extraction. Several features extraction techniques for the Speaker Identification (SI) task are used by researchers, such as Perceptual Linear Predictive (PLP), Linear Predictive Codes (LPC), Mel Frequency Cepstral Coefficients (MFCC) [14] etc. The step following the features extraction is called the modeling step, which allows the SI. Several traditional techniques are used such as modeling with the Gaussian Mixture Models (GMM), and Hidden Markov Model (HMM) [15], [16], [17], [18], [19]. These modeling methods have been compared in several works [20], [21]. Hybridization methods have been considered in this field to optimize performance [20], [21], [22], [23], [24], [25]. Classification algorithms provide better results for identification when the extracted features are distinctive and relevant, which rounds the extraction step a very important step in the SI process. Several works on SI have shown the effectiveness of PLP and MFCC. Nevertheless, these extraction techniques have identification errors when the speech signal is complex. To remedy the constraints of the mentioned techniques, we propose in this paper to use the logarithmic energies obtained directly from the energy filter bank to obtain Mel Frequency Energy Coefficients (MFECs). The MFECs coefficients are injected in the Convolutional Neural Network (CNN) model for an additional feature extraction phase, to obtain a better representation of the features. This study proposed three architectures for SI. The first two architectures used the new hybrid (MFEC-CNN) features extraction scheme. For the classification phase, one of the architectures used deep learning (CNNs). The second architecture used the model eXtreme Gradient Boosting (XGBoost) as a classifier. The third architecture used only one feature extraction method

E-mail address: d.abdiche@univ-boumerdes.dz, khaled.harrar@univ-boumerdes.dz

(MFEC) and the XGBoost for classification.
The major contributions of this work are:
1- Propose a feature extraction method that uses logarithmic energies derived directly from the filter bank for speaker identification.
2- Develop a new hybrid (MFEC-CNN) feature extraction scheme with the XGBoost classifier.
3- Compare the fit of the suggested model that employs the hybrid feature extraction approach at two levels with the one-level extraction approach.
4- Improve the SI performance compared to literature.

This article is organized as follows: Section 2 summarizes the related work.section 3 explains aspects of speaker recognition. The feature extraction and classification techniques used in this work are presented in Section 4. Section 5 explains the systems used in this paper. Section 6 reports the results obtained from the various systems used. Section 7 ends with a conclusion.

2. Related Work

Through automation and the development of services, and devices, access and control, research has accelerated to meet the needs of the industry and has been directed towards the creation of reliable and rigorous systems. The use of voice is the easiest (can be used remotely) and most reliable (voice is unique to each individual) way to meet these needs. Several works have been designed for voice-based systems such as SI systems. Soleymanpour et al. [26] used the clustering-based MFCCs technique injected into Artificial Neural Networks (ANN) classifier. They proposed two methods for obtaining the MFCCs feature vectors with the highest similarity. The accuracy rate in this study reached 93% for the ELSDSR dataset. The execution time is reduced to 20%. Sekkate et al. [27] used the technique of fusing wavelet components and Gammatone Frequency Cepstral Coefficients (GFCC) to train the Support Vector Machines (SVM) for SI. This algorithm reached a precision rate of 92.66%. In the paper of verma et al. [28] the wavelet transform was used to capture the frequency variation over time, and the MFCCs features were used to approximate the base frequency information. The K Nearest Neighbor (KNN) classifier on the Voxforge dataset made the fusion evaluation. The performance of the designed system was improved. Jahangir et al [29] presented a hybridization of MFCC and temporal coefficients (MFCCT). This hybrid method allowed a better representation of speaker-related features, which were subsequently identified with a Deep Neural Network (DNN). The model resulted in a considerable accuracy rate. Nidhyananthan et al. [30] injected Relative Spectra–Mel Frequency Cepstral Coefficients (RASTA–MFCC) features into a GMM classifier. The system used short duration data achieved 97% of accuracy. The problem is that this accuracy may not be achieved with longer duration data. In the work of Leu et al. [31] a phonetic speaker model has been established for SI based on MFCC for voice feature extraction, and the GMM model as a modeling technique. This study concluded that the SI rate is high when test voices are also used for model learning. Sekkate et al. [32] used a hybrid features extraction method using MFCC and Stationary Wavelet Transform (SWT). The classifier KNN is used for classification. The SI system in this work achieved a 92.8% rate in clean conditions and an accuracy rate of 81.80% in a noisy environment. This study has shown that using SWT rather than Discrete Wavelet Transform (DWT) in feature extraction gave better results. In the work of Zulfiqar et al. [33] speaker features have been extracted by MFCC technique. Vector Quantization (VQ) technique was used by Linde-Buzo-Gray (LBG) algorithm as a modeling technique. For the evaluation of this algorithm, two databases have been used in a noisy environment. Different manipulations in this study have shown that the accuracy rate is related to the sampling frequency and the number of VQ vectors. The identification accuracy reached 100%. This rate was obtained for the highest number of VQ vectors (64 vectors) and the highest sampling frequency 11025 Hz. Nassif et al. [34] proposed a method for SI in a noisy and emotional speech environment. They first used a noise reduction technique based on Computational Auditory Scene Analysis (CASA). Then they used a hybridization between the GMM and the CNN for emotion identification and recognition. The performance of this system was evaluated with different databases; the results of this model have been promising. The study of Abdulwahid et al. [35] aimed at identifying legal speakers. The system employed used the Arabic language and allowed noise reduction. The steps in this work consisted in applying the MFCC and VQ on the sentence of the legal speaker to obtain the feature vector which was then used for the identification using the Logistic Model Tree (LMT). This model provided an accuracy of 91.53%. KNN algorithm was also used in this work and gave an accuracy of 94.56%. In the work of Rahman et al. [36] , the data used were text independent. Static prosodic features and dynamic prosodic features were extracted and trained by a DNN for SI. Static prosodic features gave a better accuracy than dynamic prosodic features, but the combination of both gave the best accuracy rate, which was 87.72%. In the work of Shihab et al. [37], a combination of CNN and Grid Recurrent Unit (GRU) was proposed for the feature extraction step for SI in a real environment. This step has been enhanced by a feature selection technique that sorts out the most relevant features. The method provided an accuracy rate of 93.51%. In the work of Ghiurcau et al. [38] SI was evaluated under different emotional states of the speakers with the Berlin database which contains different emotional states. For this purpose, MFCC was used for feature extraction and GMM for feature vector modeling. The results were good and increased up to 98% when the system used the different emotions during the training phase. In the work of Yadav et al. [39] the wavelet transform was used for SI. Silences were removed by a preprocessing phase; DWT-based MFCC and traditional MFCC were used for feature extraction. VQ was used to compare between the obtained vectors and the reference vectors, and the LBG algorithm was used to

determine the identity of the speakers based on a certain threshold. The method adopted resulted in an accuracy rate of 85%. Although the literature has shown that MFCCs were the most suitable features for speech recognition, the results obtained so far for automatic SI were not promising. To improve the accuracy rate, this work proposed new MFEC features and a two-level feature extraction step.

3. Speaker Recognition Process

Automatic Speaker Recognition (ASR) is the process of extracting the identity of a speaker with the help of a machine from his voice signal. This task using the speech signal is robust because this signal is unique and specific to each individual. The speech signal is characterized by two types of variability, an intra-speaker variability due to medical conditions or advancing age, and an inter-speaker variability due to the physical and anatomical differences of the speech apparatus.

There are two operating modes in speaker recognition schemes, text independent and text dependent. The text-dependent mode has a higher security aspect because it limits recognition to a specific number of words, which the system must learn. The speaker-independent system can recognize a speaker with any spoken word. It is less accurate and more flexible than the speaker-dependent system. In addition, speaker recognition includes several issues, mainly identfication and verification. Speaker identification is the ability to identify a speaker statement from a group of previously defined statements in the database. The identification is a multiclass classification (Figure 1). The verification consists of checking whether the speaker's identity is what it claims to be. The verification is a binary classification (acceptance or rejection).

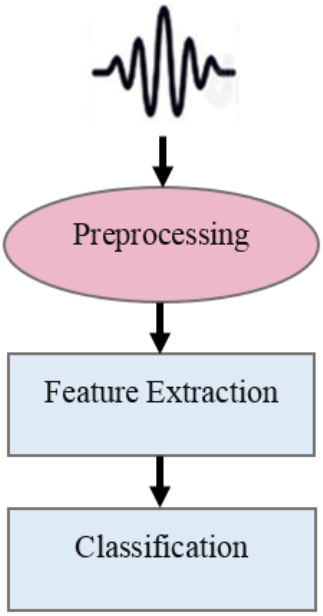

Figure 1. Automatic speaker identification process

The process followed in ASR starts with the preprocessing of the raw audio data, provided by the database. The next step is very important, it consists of feature extraction. There are several parameterization techniques [40], [41]. The feature extraction provides feature vectors that serve as input for the next phase and allow to have a good classification, and consequently a good recognition. Figure 2 resumes the speaker recognition task.

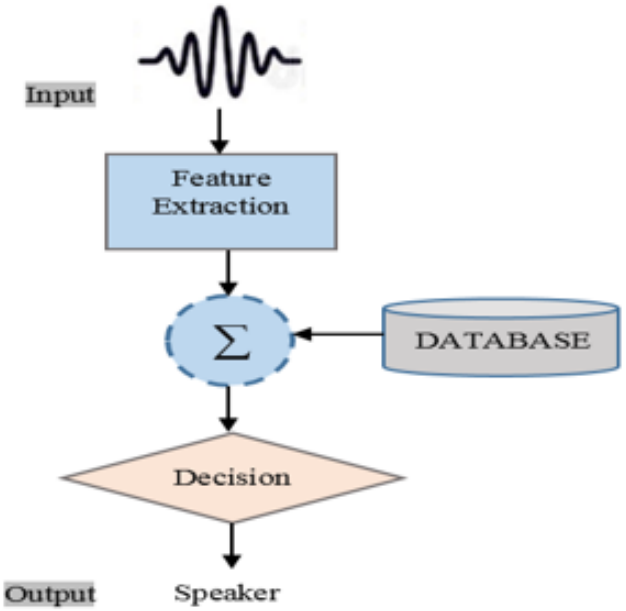

Figure 2. Speaker recognition process

In our model design, we have used the MFECs coefficients as features. This technique is presented in the following section.

4. Methodology

A. *Mel Frequency Cepstral Coefficients*

This work is based on the identification of the speaker using machine learning and deep learning techniques. A set of vocal signals of speakers with acoustic characteristics based on MFCC were used, which allows a compact representation of the data. Several studies have shown that MFCCs provided good results for ASR [42] , but also for

speaker identification [43] . The MFCCs coefficients give a very good illustration of the phoneme produced due to their efficiency in reproducing the envelope of the power spectrum, where the vocal tract appears. Obtaining the MFCCs coefficients involves a series of calculations applied to the power spectrum. The first of which will increase the energy of high frequencies and avoid false extraction of the characteristics. The first stage, known as pre-emphasis, whose equation for the applied filter is as follows [44] :

$$X(t) = z(t) - a\ z\ (t - 1) \quad (1)$$

Where a = [0.95, 0.98]

The segmentation of the signal is applied to have several signals of very short duration or they can be considered quasi-stationary [45] which means that the signal is stable. This operation leads to a distortion of the signal which leads to an operation of multiplication by weighting window to find continuous signals. In the broad field of speaker recognition, the windowing step is often done by using the Hamming window W. It minimizes spectral distortion by minimizing the ends of the frame to increase the continuity of the signal. Its equation is given as follows [46] :

$$W(n) = 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right],\ 0 \le n \le N-1 \quad (2)$$

Where N represents the number of samples. The Fast Fourier Transform (FFT) is used to convert samples to the frequency domain, as follows:

$$Z_2(n) = \sum_{k=0}^{N-1} Z_1(k)e^{-i\omega n/N} \quad (3)$$

Where, $n = 0,1,2,..., N-1$.

To obtain the spectrum, a simple calculation that consists of taking the square of the magnitude of each frequency component is performed as follows:

$$Z_3(n) = (real\ (Z_2(n)))^2 + (imag\ (Z_2(n)))^2 \quad (4)$$

The spectrum smoothing must be performed using a series of filters called filter bank to keep only the envelope of the spectrum to which we are interesting, and also to decrease the dimensions of the spectral vectors. This group of filters is a succession of band-pass filters of triangular shape aiming to increase the energy of low frequency signals and to decrease that of high frequencies, so that the signals match the evolutions of the Mel scale. The latter has a strong coherence with the frequency scale of the human ear, its equation is given by [47] :

$$F_{MEL} = 2595 \times log_{10}(1 + f_{Hz}/700) \quad (5)$$

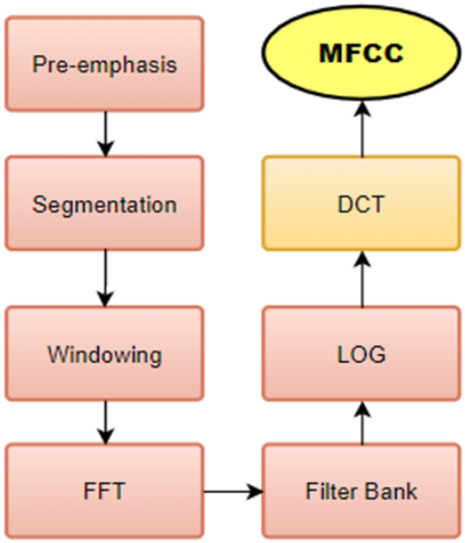

Figure 3. MFCC feature estimation

The logarithm of each previous value is calculated as follows:

$$Z_5(n),\ = ln\ (Z_4(n),)\ où\ 0 \le n < k. \quad (6)$$

The last step of this technique is the application of the Discrete Cosine Transform (DCT) which allows returning to the time domain :

$$Z_6(n) = \sum_{k=1}^{k} Z_5(k)\cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{k}\right],\ n = 1,2,\ 3,\ \ldots,\ k \quad (7)$$

The result of this DCT is the so-called the MFCC coefficients. The components that carry little information related to the speaker are excluded such as the average value of the input signal, and thus only the highly representative vectors are used [48] . The calculation of the MFEC coefficients is similar to the calculation of the MFCC coefficients, but without the last operation which is the calculation of the DCT. The flowchart of this calculation is presented in Figure 3.

B. Convolutional Neural Network

The human was strongly inspired by the world around him for the design of various algorithms, such as machine learning. Deep learning based on the behavior of human neurons is part of machine learning. In this subset belong CNN. The connection between neurons in such a network is inspired by the visual cortex of animals. CNNs were first presented by Yan Le et al. [49] . In this last decade, CNNs have seen success due to the results they achieved for recognition and classification of images. CNN is a DNN

with a double function, it allows in a first step the extraction of features and in a second step, the classification (Figure 4). CNNs use the concept of weight sharing to decrease the number of parameters in the training phase, which leads to a reduction in the memory space consumed and therefore in the computation time [50] . The concept of translational invariance allows CNN to reveal the class membership of the input, despite the translations that may occur in the network, thus reducing the complexity of its structure. A typical CNN network is composed mainly of convolution filters to which the weights are adjusted during the learning phase by the backpropagation algorithm [51] . This algorithm minimizes the output error compared to the training values. The convolution is an addition/multiplication sequence, which results in linear output values. Thus, a ReLU function written as: *R(y) = max (0, y)*, overcomes this problem by suppressing negative values, which forces the neurons to return positive values. Generally, after each one or two convolution layers, there is a pooling layer employed. It helps to minimize the fitting and keep the model simple. Several types of pooling can be employed such as maximum clustering or average clustering. This step allows to replace several values by one, and thus pooling also participates in reducing the size of the network. Fully Connected (FC) layers are Multi-Layer Perceptrons (MLP), i.e., Neural Networks (NN) containing several hidden layers, called computational layers, as well as input and output layers. The FC layers perform a weighted sum of the entry values, each with its weight. The FCs apply an activation function. Generally, for a multiclass task like the SI task, the activation function used is the Softmax function which is given by Equation 8:

$$softmax\left(z_j\right) = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}}, j = 1, 2, \ldots \tag{8}$$

Where Z_i represents the output of j, e_j^Z is the exponential value of Z_j and k is the constituent of vector Z

C. EXtreme Gradient Boosting

EXtreme Gradient Boosting [52] or briefly called XGBoost is an algorithm that belongs to the family of ensemble classifiers. This algorithm brings higher performance and speed. For this purpose, it is advisable to use decision trees (weak), which are boosted by the gradient. The type of decision tree in such algorithm is a group of Classiffication And Regression Trees (CARTs). These trees are added sequentially by applying the adaptive strategy, which allows for corrections and minimization of the loss function [53]. This repetitive process stops only when the objective function finds that no further improvement can be brought, making this algorithm sensitive to overfitting. On the other hand, the XGBoost model performs the missing values processing as well as the parallel processing. During the learning phase, this algorithm presents, mainly, efficiency of resources in time and memory. Currently, it is one of the most robust recognition algorithms. This algorithm provided high accuracy rates in various works related to speech classification [54], [55]. Given supervised learning in XGBoost algorithm, the model that allows the prediction of target variable y_i by input variablesx_i is given by:

$$\widehat{Y_i} = \sum_{N=1}^{N} s_N\left(z_i\right), s_N \in S \tag{9}$$

With:

N represents the number of trees.

f_N refers to the function of the N_{th} tree in S.

S represents the functional space of the tree.

During training, each tree trained has the role of completing the residual so far, hence XGBoost minimizes the following regularized objective:

$$L = \sum_{i=1}^{n} l(y_i, \widehat{y_i^t}) + \sum_{i=1}^{t} \Omega\left(f_i\right), \tag{10}$$

L refers to the loss function. In addition, Ω is the regularized term.

With:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{11}$$

Where T is the number of leaves, and w_j represents the score of the j_{th} leaf. When the regularization term γ has reached the optimum state, the gradient descent is used for different loss functions.

5. Proposed Method

The proposed method uses the logarithmic energies obtained directly from the energy filter bank to obtain the MFECs. The MFECs coefficients are injected into the CNN network for another features extraction step to get a better illustration of the features. This study proposed three architectures for speaker identification. The first two architectures used the new hybrid features extraction scheme (MFEC-CNN). For the classification phase, one of the architectures used deep learning, specifically CNNs, and the other architecture used XGBoost as a classifier. The third architecture used a single feature extraction method (MFEC) and XGBoost for classification. This one allows to see the importance of good feature extraction using the hybrid method. The proposed architectures are presented in Figure 5. The hyperparameters used in the proposed CNN are given in Table I.

The proposed CNN architecture consists of seven convolution layers and four pooling layers, followed by two fully

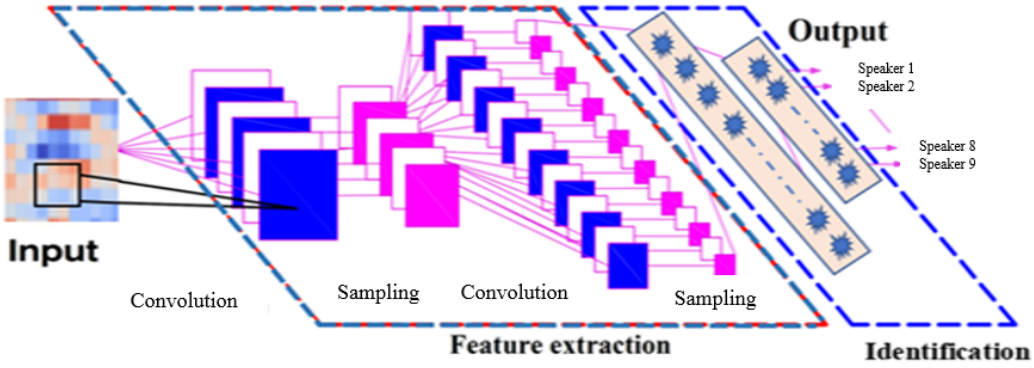

Figure 4. Convolutional neural network

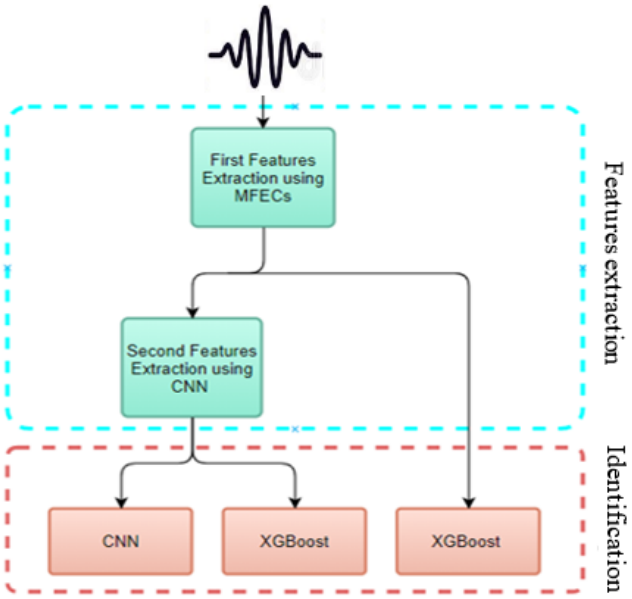

Figure 5. Identification model

connected layers. All convolution layers used filters of size (3×3) with stride = 1. The first three convolution layers, the fourth convolution layer, and the last three convolution layers include 16, 32, and 64 filters respectively. To get robust learning and avoid overfitting, we have used several small filters [56], [57]. Filters of size 2×2 with a stride = 2 were used in the pooling layers, and were succeeded by a normalization batch layer. The dropout is used to prevent the overfitting of the data. The last part of the network is composed of two fully connected layers, and a normalization batch layer. The Softmax activation function was used after the last fully connected layer. The use of CNNs involves a set of convolution operations applied to particular regions in an image (CNNs have locality property). The research works that include speaker recognition have shown that the MFCC coefficients give the best results in the state of art, compared to other feature extraction techniques [40], [41], [42]. This is why our choice orientation is towards MFCCs. Nevertheless, these coefficients have a drawback when computing the DCT, which is the loss of locality property. To improve the performance of identification in this direction, and to benefit from the high recognition rates that MFCCs and CNNs have gained, it is necessary to go beyond the concern for the locality.

The resulting coefficients are called MFEC coefficients and were us as input to our three networks: (MFEC-CNN)-CNN, (MFEC-CNN)-XGBoost. and MFEC-XGBoost. XGBoost is based on gradient boosting that reduces errors in sequence models. It can be used for multiclass classification. The parameters of this model are given in Table II.

6. Results and discussion

A. Dataset

To validate the system, the Voxforge database was used [58]. This database is an open library of transcribed multilingual speech patterns. Voxforge collects speech signals from speakers. Voxforge integrates researchers and human voice donors from different parts of the world, where any registered person can send his voice recording from a microphone and facilitate the study of the human voice. The model consisted of 70 randomly selected speakers, with 2115 samples of which 1692 samples were used for training the system and 423 for testing. This represents 80% and 20% for training and testing respectively. Each speaker reads sentences in English that are registered at a sampling frequency of 8 kHz. The speech samples are in a time interval of 2 - 10 seconds. The format adapted to the files is the wav format. Since the system of identification is text independent, each speech is different from the other.

TABLE I. HYPERPARAMETERS USED FOR THE CNN ARCHITECTURE

Layer	Input size	Output size	Stride/Padding	Parameters
Conv	32×112×1	32×112×16	S=1/Pad=1	160
Conv	32×112×16	32×112×16	S=1/Pad=1	2320
Maxpool	32×112×16	16×56×16	S=2	0
Conv	16×56×16	16×56×16	S=1/Pad=1	2320
Maxpool	16×56×16	8×28×16	S=2	0
Conv	8×28×16	8×28×32	S=1/Pad=1	4640
Conv	8×28×32	8×28×64	S=1/Pad=1	18496
Maxpool	8×28×64	4×14×64	S=2	0
Conv	4×14×64	4×14×64	S=1/Pad=1	36928
Conv	4×14×64	4×14×64	S=1/Pad=1	36928
Maxpool	4×14×64	2×7×64	S=2	0

TABLE II. XGBOOST PARAMETERS

Parameters	Parameters values
Estimators	125
Learning rate	0.06
Maximum depth of the tree	4
Minimum child weight	1
Gamma	0
Reg-lambda	1
Reg-alpha	0
Objective	multi:softprob

B. Experimental results

In this work, we have performed three architectures for identification. We consider:

- In a first architecture: a double feature extraction using MFEC and CNN as feature extraction and classification method.
- In a second architecture: a double feature extraction using the MFEC coefficients and CNN only as a characteristic extraction technique, and the XGBoost as a classifier.
- In a third architecture: a single feature extraction using MFECs and classification using XGBoost. The same database, the same MFECs coefficients, the same CNN architecture, and the same XGBoost classifier were used in each of the architectures to better assess the representation of the higher-level features. The accuracy factor was utilized to measure the efficiency of the three architectures. A comparison study with literature methods operating with the same database (Voxforge) was established to demonstrate the performance of our network. The results obtained for the (MFEC-CNN)/CNN architecture are given in Figure 6 and Figure 7 and provide respectively the accuracy and loss rate for SI.The accuracy rates for the three proposed methods are given in Figure 8.

In this work, four statistical metrics (Accuracy, Recall, Precision, and F1 score) were used for the evaluation of the performance of our system for the three proposed

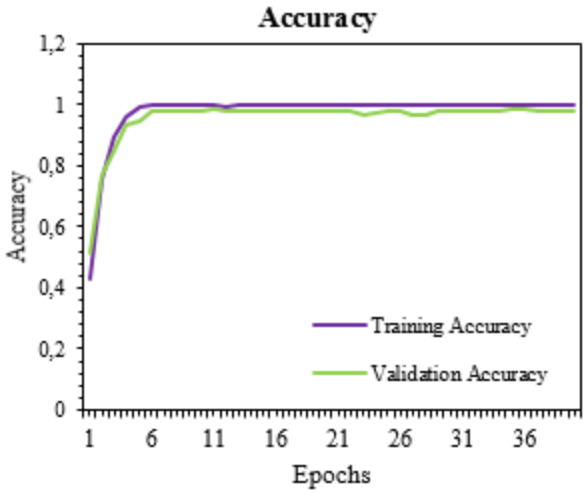

Figure 6. Accuracy rate obtained for speaker identification

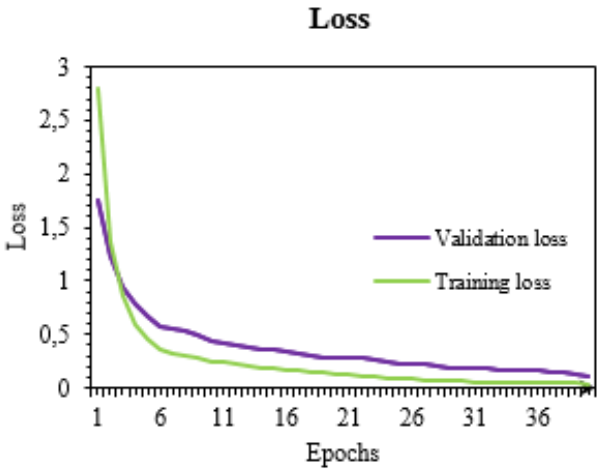

Figure 7. Loss rate obtained for speaker identification

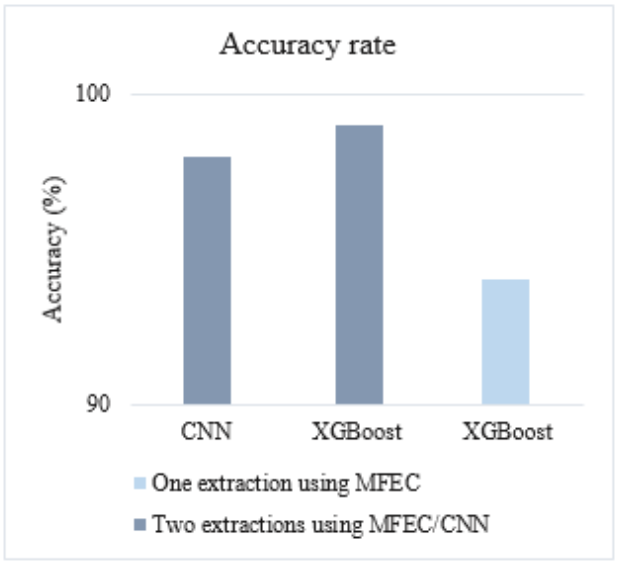

Figure 8. Accuracy rate for proposed methods

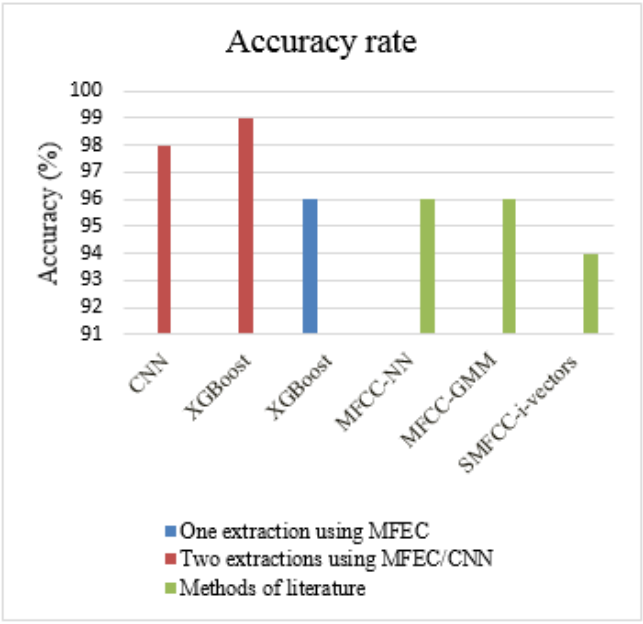

Figure 9. Accuracy rate for the comparative study

methods. The different values of these parameters are given in Table III.

A comparison with the literature approaches using the Voxforge database for SI is established and presented in Table IV and Figure 9. For the MFCC-NN method [59], the authors used a NN on 15 MFCC coefficients. The Voxforge database was used with 10 speakers in the first experiment and 20 speakers in a second experiment. In the work of Alfredo Maesa et al. [60], 20 MFCC coefficients were used and statistically analyzed by GMM with 32 GMM mixtures, and a set of speakers were randomly selected from the Voxforge dataset. In the work of Sara Sekkate et al. [61] the Sparse MFCC (SMFCC) coefficients were used for feature extraction. Modeling was done through the i-vectorial technique, and SVMs were used for classification.

C. Discussion

Table III summarizes the different parameters used for the evaluation of the three proposed methods. For the first architecture, which uses two levels of feature extraction and CNN for identification, the precision was 98.60% and the Recall was 97%, while for the second architecture which also uses two levels of feature extraction but with an XGBoost classifier, the precision achieved 100% and the Recall reached 99%. This leads to conclude that the SI system that uses two levels of extraction with an XGBoost algorithm for identification is much better than the identification with a CNN. For the third architecture, which uses a single level of feature extraction with XGBoost algorithm for identification, the Recall was 94% with an F1 score of 0.954. These results show the effectiveness of the proposed hybrid method that uses two levels of feature extraction. This allows a higher-level feature representation.

Table IV presents a comparison that summarizes the accuracy rates of the literature studies and the proposed methods for ASI (text-independent speech) taken from the Voxforge open-source database. It shows that the two proposed methods that use MFEC-CNNs as extractors have the highest accuracy rates. The architecture using XGBoost reached a higher rate (99.45%) than the architecture using CNNs as a classifier (98.02%). The third architecture using MFEC/XGBoost achieved a rate of 96.8%. The results of the methods existing in the literature were lower than the results obtained by our proposed methods. The methods using handcrafted modeling techniques such as SMFCC/i-vectors and MFCC-GMM present respectively an accuracy rate of 94.29% and 96%. The MFCC-NN method provided an accuracy rate not exceeding 96%.

We have examined the performance of our system in terms of applying new MFEC features and different degrees of feature extraction. MFCCs coefficients have shown their robustness in speaker identification systems. However, these coefficients have a drawback due to the computation of DCT, which induces a loss of the locality property. To overcome this problem, we estimated the MFCCs without using DCT, which provided the MFECs coefficients that give a better illustration of the features compared to the standard MFCC method. The MFECs coefficients were then injected into the CNN network to perform a second extraction step (the CNN only takes into account the most relevant features), and obtain a better illustration of the features with less information loss. Therefore, using this higher level input data representation with an XGBoost

TABLE III. THE METRICS USED TO EVALUATE THE PERFORMANCE OF THE SYSTEM.

Algorithm	Accuracy (%)	Recall (%)	Precision (%)	F1 score
(MFEC-CNN)-CNN	98.02	97	98.60	0.977
(MFEC-CNN)-XGBoost	99.45	99	100	0.994
MFEC-XGBoost	96.80	94	97	0.954

TABLE IV. ACCURACY RATE FOR THE COMPARATIVE STUDY

Methods	Accuracy (%)
SMFCC/ I-VECTOR [61]	94.29
MFCC/GMM (32 Gaussians) [60]	96.0
MFCC/NN [59]	96.0
MFEC/XGBoost (Proposed)	96.80
MFEC/CNN (Proposed)	98.02
MFEC-CNN /XGBoost (Proposed)	99.45

classifier gave the best accuracy rate, which is 99.45%, and a precision rate that is 100%. This new method offered an improvement of almost 6% over the state of the art methods using MFCCs only (Table IV). The method using only MFECs did improve the performance of the system by almost 3% over the methods using MFCCs (Table III). However, the use of the hybrid method (MFEC-CNN) offers twice this improvement, and the text-independent SI rates obtained by this method are the highest.

7. Conclusion

This paper aimed to show the importance of having a high-level feature representation to improve speaker identification performance. However, the use of the CNN extractor and logarithmic energies obtained from the filter bank energies (MFEC) instead of the raw wave signal resulted in higher identification rates compared to the method where only a single feature extraction is used. Moreover, the combination of the CNN-XGBoost improved the performance of identification. The comparative study with the literature approaches demonstrated the efficiency of our model and the importance of the extraction phase where a high accuracy rate was achieved (99.45%). Increasing the iteration level of the CNN optimization program and using a deeper architecture as well as increasing the number of XGBoost estimators provide higher accuracy of classification. We suggest in future work to use a deeper CNN architecture, and a fusion of several classifiers.

References

[1] H. Beigi, "Speaker recognition," in *Fundamentals of Speaker Recognition*. Springer, Boston, MA, 2011, pp. 543–559.

[2] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, 1996.

[3] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B.-H. Juang, "A vector quantization approach to speaker recognition," *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 10, pp. 387–390, 1985.

[4] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 1 2000.

[5] S. Furui, "Recent advances in speaker recognition," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 859–872, 1997, audio- and Video-Based Person Authentication. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865597000731

[6] S. Pruzansky, "Pattern[U+2010]matching procedure for automatic talker recognition," *The Journal of the Acoustical Society of America*, vol. 35, 1963.

[7] P. D. Bricker and S. Pruzansky, "Effects of stimulus content and duration on talker identification," *The Journal of the Acoustical Society of America*, vol. 40, pp. 1441–1449, 12 1966.

[8] F. Garzia, E. Sammarco, and R. Cusani, "The integrated security system of the vati can city state," *International Journal of Safety and Security Engineering*, vol. 1, 2011.

[9] G. Contardi, F. Garzia, and R. Cusani, "The integrated security system of the senate of the italian republic," 7 2011, pp. 103–115.

[10] F. Garzia, E. Sammarco, and R. Cusani, "Vehicle/people access control system for security management in ports," *International Journal of Safety and Security Engineering*, vol. 2, pp. 351–367, 12 2012.

[11] F. Garzia and R. Cusani, "The integrated safety/security/communication system of the gran sasso mountain in italy," *International Journal of Safety and Security Engineering*, vol. 2, pp. 13–39, 3 2012.

[12] J. Naik and G. Doddington, "Evaluation of a high performance speaker verification system for access control," in *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, 1987, pp. 2392–2395.

[13] M. G. Gomar, "System and method for speaker recognition on mobile devices," May 26 2015, uS Patent 9,042,867.

[14] A. Maurya, D. Kumar, and R. Agarwal, "Speaker recognition for hindi speech signal using mfcc-gmm approach," *Procedia Computer Science*, vol. 125, pp. 880–887, 2018, the 6th International Conference on Smart Computing and Communications. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050917328806

[15] T. Mahboob, M. Khanam, M. Khiyal, and R. Bibi, "Speaker identification using gmm with mfcc," *International Journal of Computer Science Issues*, vol. 12, pp. 126–135, 03 2015.

[16] D.-P. Munteanu and S.-A. Toma, "Automatic speaker verification experiments using hmm," in *2010 8th International Conference on Communications*, 2010, pp. 107–110.

[17] P. Varchol, D. Levicky, and J. Juhar, "Optimalization of gmm for text independent speaker verification system," in *2008 18th International Conference Radioelektronika*, 2008, pp. 1–4.

[18] M. S. Sinith, A. Salim, K. G. Sankar, K. V. S. Narayanan, and V. Soman, "A novel method for text-independent speaker identification using mfcc and gmm," in *2010 International Conference on Audio, Language and Image Processing*, 2010, pp. 292–296.

[19] S. J. Abdallah, I. M. Osman, and M. E. Mustafa, "Text-independent speaker identification using hidden markov model," 2012.

[20] I. Naseem, R. Togneri, and M. Bennamoun, "Sparse representation for speaker identification," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 4460–4463.

[21] Y. Liu, L. He, Y. Tian, Z. Chen, J. Liu, and M. T. Johnson, "Comparison of multiple features and modeling methods for text-dependent speaker verification," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 629–636.

[22] A. Rashno, S. M. Ahadi, and M. Kelarestaghi, "Text-independent speaker verification with ant colony optimization feature selection and support vector machine," in *2015 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA)*. IEEE, 2015, pp. 1–5.

[23] J. S. Devi, S. Yarramalle, S. P. Nandyala, and P. V. B. Reddy, "Optimization of feature subset using habc for automatic speaker verification," in *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. IEEE, 2017, pp. 1–6.

[24] J. S. Devi and S. Yarramalle, "Multi objective optimization problem resolution based on hybrid ant-bee colony for text independent speaker verification," *IJ Modern Education and Computer Science*, vol. 1, pp. 55–63, 2015.

[25] R. Djemili, M. Bedda, and H. Bourouba, "A hybrid gmm/svm system for text independent speaker identification," *International Journal of Computer and Information Science & Engineering*, vol. 1, no. 1, 2007.

[26] M. Soleymanpour and H. Marvi, "Text-independent speaker identification based on selection of the most similar feature vectors," *International Journal of Speech Technology*, vol. 20, no. 1, pp. 99–108, 2017.

[27] S. Sekkate, M. Khalil, and A. Adib, "A feature level fusion scheme for robust speaker identification," in *International Conference on Big Data, Cloud and Applications*. Springer, 2018, pp. 289–300.

[28] G. K. Verma, "Multi-feature fusion for closed set text independent speaker identification," in *International conference on information intelligence, systems, technology and management*. Springer, 2011, pp. 170–179.

[29] R. Jahangir, Y. W. Teh, N. A. Memon, G. Mujtaba, M. Zareei, U. Ishtiaq, M. Z. Akhtar, and I. Ali, "Text-independent speaker identification through feature fusion and deep neural network," *IEEE Access*, vol. 8, pp. 32 187–32 202, 2020.

[30] S. S. Nidhyananthan, R. S. S. Kumari, and T. S. Selvi, "Noise robust speaker identification using rasta–mfcc feature with quadrilateral filter bank structure," *Wireless Personal Communications*, vol. 91, no. 3, pp. 1321–1333, 2016.

[31] F.-Y. Leu and G.-L. Lin, "An mfcc-based speaker identification system," in *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*. IEEE, 2017, pp. 1055–1062.

[32] S. Sekkate, M. Khalil, and A. Adib, "Fusing wavelet and short-term features for speaker identification in noisy environment," in *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*. IEEE, 2018, pp. 1–8.

[33] A. Zulfiqar, A. Muhammad, and M. E. AM, "A speaker identification system using mfcc features with vq technique," in *2009 Third International Symposium on Intelligent Information Technology Application*, vol. 3. IEEE, 2009, pp. 115–118.

[34] A. B. Nassif, I. Shahin, S. Hamsa, N. Nemmour, and K. Hirose, "Casa-based speaker identification using cascaded gmm-cnn classifier in noisy and emotional talking conditions," *Applied Soft Computing*, vol. 103, p. 107141, 2021.

[35] S. Abdulwahid, M. A. Mahmoud, and N. Abdulwahid, "Arabic speaker identification system for forensic authentication using k-nn algorithm," in *International Visual Informatics Conference*. Springer, 2021, pp. 459–468.

[36] A. Rahman and W. C. Wibowo, "Deep neural network for speaker identification using static and dynamic prosodic feature for spontaneous and dictated data," *JISIP (Jurnal Ilmu Sosial dan Pendidikan)*, vol. 5, no. 4, 2021.

[37] M. S. H. Shihab, S. Aditya, J. H. Setu, K. Imtiaz-Ud-Din, and M. I. A. Efat, "A hybrid gru-cnn feature extraction technique for speaker identification," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2020, pp. 1–6.

[38] M. V. Ghiurcau, C. Rusu, and J. Astola, "A study of the effect of emotional state upon text-independent speaker identification," in *2011 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 4944–4947.

[39] S. S. Yadav and D. Bhalke, "Speaker identification system using wavelet transform and vq modeling technique," *International Journal of Computer Applications*, vol. 112, no. 9, 2015.

[40] E. Yücesoy and V. V. Nabiyev, "Comparison of mfcc, lpcc and plp features for the determination of a speaker's gender," in *2014 22nd Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2014, pp. 321–324.

[41] N. Dave, "Feature extraction methods lpc, plp and mfcc in speech recognition," *International journal for advance research in engineering and technology*, vol. 1, no. 6, pp. 1–4, 2013.

[42] C. Hanilci and F. Ertas, "Vq-ubm based speaker verification through dimension reduction using local pca," in *2011 19th European Signal Processing Conference*. IEEE, 2011, pp. 1303–1306.

[43] C. Kumar, F. Ur Rehman, S. Kumar, A. Mehmood, and G. Shabir, "Analysis of mfcc and bfcc in a speaker identification system," in *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. IEEE, 2018, pp. 1–5.

[44] C. Turner, A. Joseph, M. Aksu, and H. Langdond, "The wavelet and fourier transforms in feature extraction for text-dependent, filterbank-based speaker recognition," *Procedia Computer Science*, vol. 6, pp. 124–129, 2011.

[45] B. H. Juang and T. Chen, "The past, present, and future of speech processing," *IEEE signal processing magazine*, vol. 15, no. 3, pp. 24–48, 1998.

[46] A. Goel and A. Gupta, "Design of satellite payload filter emulator using hamming window," in *2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom)*. IEEE, 2014, pp. 202–205.

[47] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Transactions on speech and audio processing*, vol. 7, no. 5, pp. 525–532, 1999.

[48] T. D. H. Kekre and V. Kulkarni, "Speaker identification using row mean of dct and walsh hadamard," *International Journal on Computer Science and Engineering (IJCSE), ISSN*, pp. 0975–3397.

[49] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

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

[51] D. Zipser and R. A. Andersen, "A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons," *Nature*, vol. 331, no. 6158, pp. 679–684, 1988.

[52] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[53] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[54] J. V. Egas-López and G. Gosztolya, "Predicting a cold from speech using fisher vectors; svm and xgboost as classifiers," in *International Conference on Speech and Computer*. Springer, 2020, pp. 145–155.

[55] J.-M. Long, Z.-F. Yan, Y.-L. Shen, W.-J. Liu, and Q.-Y. Wei, "Detection of epilepsy using mfcc-based feature and xgboost," in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2018, pp. 1–4.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[57] A. Torfi, J. Dawson, and N. M. Nasrabadi, "Text-independent speaker verification using 3d convolutional neural networks," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[58] "Voxforge," http://www.voxforge.org.

[59] Z. Kozhirbayev, B. A. Erol, A. Sharipbay, and M. Jamshidi, "Speaker recognition for robotic control via an iot device," in *2018 World Automation Congress (WAC)*. IEEE, 2018, pp. 1–5.

[60] A. Maesa, F. Garzia, M. Scarpiniti, R. Cusani *et al.*, "Text independent automatic speaker recognition system using mel-frequency cepstrum coefficient and gaussian mixture models," *Journal of Information Security*, vol. 3, no. 04, p. 335, 2012.

[61] S. Sekkate, M. Khalil, and A. Adib, "Speaker identification for ofdm-based aeronautical communication system," *Circuits, Systems, and Signal Processing*, vol. 38, no. 8, pp. 3743–3761, 2019.

Dehia ABDICHE is preparing her Ph.D. at the University of M'Hamed Bougara, Boumerdes. In 2019, she obtained her Master degree in Telecommunications Systems at the University of Houari Boumédienne, Algiers. Her research interests include artificial intelligence, computer vision, deep learning, signal and image processing, machine learning, including texture characterization.

Khaled Harrar received his Ph.D. degree in 2014 in Electronics from National Polytechnic School (ENP), Algiers. He is a member of the LIST laboratory (Boumerdes University), and also a reviewer in several conferences and international journals (Elsevier, Springer, etc.). Since 2004, he is an Associate Professor at Boumerdes University. He supervised more than 30 Engineer and Master students in the area of Electronics, signal image processing. Currently, he is supervising 03 Ph.D. students. His research interest concerns artificial intelligence, computer vision, deep learning, signal and image processing, including medical imaging, biometrics and complex texture characterizing by fractal analysis, fractional Brownian motion models, for machine learning, computer-aided detection, and diagnosis in medical applications. With great commitment and professionalism, Khaled has been crowned by many rewards (Best paper presentation award (Madrid Spain), honor certificates, IEEE ISBI World challenge (Beijing, China), etc). He has made valuable contributions to research and has a number of publications to his credit in International Journals of high repute.