International Journal of Speech Technology (2022) 25:443–455
https://doi.org/10.1007/s10772-022-09971-y

Enhancement of spoken digits recognition for under-resourced languages: case of Algerian and Moroccan dialects

Khaled Lounnas[1] · Mourad Abbas[2,3] · Mohamed Lichouri[3] · Mohamed Hamidi[4,5] · Hassan Satori[5] · Hocine Teffahi[1]

Received: 10 March 2021 / Accepted: 27 March 2022 / Published online: 15 April 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
In this paper, we present a set of experiments aiming to improve the recognition of spoken digits for under-resourced dialects of the Maghrebi region, using a hybrid system. Indeed, integrating a Dialect Identification module into an Automatic Speech Recognition (ASR) system has shown its efficiency in previous works. In order to make the ASR system able to recognize digits spoken in different dialects, we trained our hybrid system on Moroccan Berber Dialect "MBD," Moroccan Arabic Dialect "MAD," and Algerian Arabic dialect "AAD," in addition to Modern Standard Arabic. We have investigated five machine learning based classifiers and two deep learning models: the first one is based on Convolutional Neural Network (CNN), and the second one uses two pre-trained models: Residual Deep Neural Network (Resnet50 and Resnet101). The findings show that the CNN model outperforms the other proposed methods and consequently enhances the performance of spoken digit recognition system by 20% for both Algerian and Moroccan dialects.

Keywords Dialect identification · Digits speech recognition · Algerian · Moroccan · Berber dialect · Resnet50 · Resnet101 · CNN

1 Introduction

The necessity to build an ASR able to recognize multi-dialectal speech becomes more and more important. One of the solutions to achieve such a task is to determine the dialect of the input speech. Nevertheless, identification of spoken Arabic dialects is a challenging task, particularly for fine-grained ones. This is due, on one hand, to the presence of similarity between these dialects in terms of phonological, morphological, lexical, and syntactical levels, and on the other hand, to the lack of corpora related to those vernaculars. In order to evaluate our approach, we used a corpus composed of ten digits spoken in different Algerian and Moroccan dialects, namely, Moroccan Berber Dialect, Moroccan Arabic Dialect, and Algerian Arabic dialect, in addition to Modern Standard Arabic. This corpus is recorded by twenty four speakers, ten times, in the three aforementioned dialects and MSA. We prepared this dataset for building models for both dialect identification and ASR systems. The work presented in this paper is twofold: first, performing Maghrebi dialects identification, and second, showing its impact on multi-dialect ASR accuracy. Our approach of identifying the dialects is based on a multitude of efficient classification algorithms, namely: k-Nearest Neighbours

✉ Khaled Lounnas
klounnas@usthb.dz; lounnaskhaled912@gmail.com

Mourad Abbas
m_abbas04@yahoo.fr

Mohamed Lichouri
m.lichouri@crstdla.dz

Mohamed Hamidi
mohamed.hamidi.5@gmail.com

Hassan Satori
hsatori@yahoo.com

Hocine Teffahi
hteffahi@usthb.dz

1 Laboratory of Spoken Communication and Signal Processing, USTHB, Algiers, Algeria

2 High Council for the Arabic Language, HCLA, Algiers, Algeria

3 Computational Linguistics Department, CRSTDLA, Algiers, Algeria

4 Multimedia and Arts department, FLLA, UIT, Kenitra, Morocco

5 LISAC, Department of Mathematics and Computer Science, FSDM, USMBA, Fes, Morocco

(KNN), and Extratrees (EXT), and Random Forest (RF), and Gradient Boosting (GB), and Convolutional Neural Networks (CNN), and Support Vector Machine (SVM) (Campbell et al., 2006). This paper is organized as follows: we present an overview of both speech-based dialect identification and recognition of dialectal speech, and the related work in Sects. 2 and 3, respectively. In Sect. 4, we describe the corpus used to run different experiments. In Sect. 5, we present the system architecture. Section 6 is devoted for experiments and results regarding both dialect identification and speech recognition. The conclusion is presented in Sect. 7.

2 Speech based dialect identification

Speech-based dialect identification attracted the interest of many researchers (Liu & Hansen, 2011; Chittaragi et al., 2018, 2019; Kakouros et al., 2020). However, there is a very few little research devoted for Arabic dialects. To supply more resources for Arabic and its dialects (Shon et al., 2020) provided a huge dialectal Arabic corpora containing 17 dialects. For this purpose, a total of 3000 h of speech were available for training a fine-grained Arabic dialects identification system, split into three subsets according to their durations (< 5 s, 5 s ~ 20 s and > 20 s). Further, many state-of-the-art techniques were built using the aforementioned corpus. The obtained results show that the longer the duration of the utterance (in this case > 20 s), the better its identification. Regarding the same problem and to highlight the usefulness of the X-Vector technique on Arabic spoken dialect identification task, (Hanani & Naser, 2020) designed an X-Vector model using a set of relevant features (acoustic, lexical, and phonetic) extracted from VarDial 2018 and VarDial 2017 and showed that it outperforms other state-of-the-Art models, for instance, those based on i-vectors, Bottleneck features, and GMM-tokens.

In the case of Maghrebi dialects, (Lounnas et al., 2018) carried out a set of experiments using different features configurations to discriminate between Standard Arabic and one of the Berber dialects known as Kabyl[1]. They showed that the combination of acoustic (Mel Frequency Cepstral Coefficients) and prosodic (melody and stress) characteristics are the appropriate representation to identify these dialects. A further extension of this work is the one developed in Lounnas et al. (2019) where different systems have been built for the purpose of identifying Persian, German, English, Arabic, and Kabyl dialect. The results showed that despite the small size of data, the system yielded an encouraging accuracy of 84.6%. Prosodic information characterized by rhythm and intonation has been used in Bougrine et al. (2018) to model six Algerian dialects, using SVM based on the Universal Pearson VII Kernel function (PUK). The authors found that prosodic cue was suitable even with a short duration of utterances with a precision of more than 69%.

In Belgacem et al. (2010), the authors have developed a GMM-based model that detects similarities between nine dialects. They showed that there are no clear borders between dialects as well as the system's ability to distinguish between eastern and western dialects and between Gulf and North African dialects, resulting in an accuracy of 73.33%. A similar approach has been presented in Nour-Eddine and Abdelkader (2015), Lachachi and Adla (2016), addressed the problem of Minimal Enclosing Ball reduction using two systems based on SVM; both are used for data reduction. These techniques were evaluated on a Maghrebi database containing five dialects (3 Algerian, 1 Moroccan, 1 Tunisian). In Terbeh et al. (2018), the authors proposed a statistical approach based on the phonetic modelling to identify the corresponding Arabic dialect for each input acoustic signal by calculating the appropriate phonetic model; then, they compared this latter to all referenced Arabic dialect models using cosine similarity.

3 Speech recognition for dialects

Many works have been tackled for recognizing Arabic Spoken Digits (Wazir & Chuah, 2019; Azim et al., 2021; Touazi & Debyeche, 2017; Zerari et al., 2018). Unfortunately, there is little research that has been done for dialectal Maghrebi speech recognition. In Satori and ElHaoussi (2014), the authors addressed the problem of speech recognition for one specific Moroccan dialect, "Tarifit Berber." They developed an ASR system for this vernacular using the CMU-sphinx tool. Sixty native speakers of Tarifit Berber have recorded a corpus composed of 10 digits and 33 alphabets. The findings showed that a 16-GMM system provided a good recognition rate of 92%. Furthermore, in order to check the ability of the HMM speech recognition system to distinguish the vocal print of Moroccan dialect speakers, it has been shown in Mouaz et al. (2019) that using MFCC, delta, and delta-delta for dialectal model design is enough for a good characterization of Moroccan dialect, yielding an accuracy of 90%. In a similar way, in El Ghazi et al. (2011), authors presented their ASR system for Moroccan dialect where they showed that HMM outperformed the dynamic programming with an accuracy of 30%.

4 Dataset preparation

Our main goal is to present the best dialect identification system which improves multi-dialect ASR performance. The lack of labelled data and standardized orthography for

[1] Kabyl is an Algerian Berber dialect.

Table 1 The corpus' characteristics

Sampling rate	16 Khz
Number of bits	16 bits
Number of channels	1, Mono
Audio data file format	.wav
# speakers	24
# tokens per speaker	100
# speakers according to gender	12 males and 12 females
Total number of tokens	2400
Number of digits	10 digits (MSA)
	10 digits (MBD)
	10 digits (AAD)
	10 digits (MAD)
Number of repetitions per word	10
Condition of noise	normal life
Preemphased	$1 - 0.97z^{-}1$
Window type hamming	25.6 ms
Frames overlap	10 ms

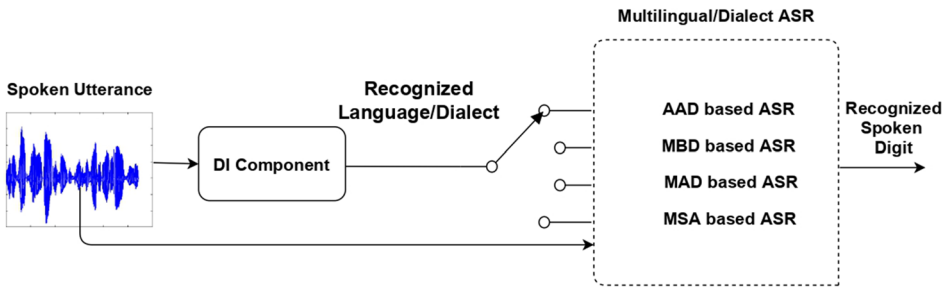
5 System architecture

Our system is based on two components: the Dialect Identification (DI) and the Automatic Speech Recognition (ASR). Figure 1 presents an illustration of the proposed architecture. The DI block aims at identifying the dialect/language of the spoken digits. This output is very important because it allows selecting the appropriate model corresponding to the dialect of the spoken utterance.

5.1 Dialect identification component

To boost our system to better recognize spoken digits, it is essential to set up a language model adaptation process. This can be done by implementing a module that identifies the dialect of spoken digits.For the sake of implementing a reliable dialect identification module, we proposed two

Fig. 1 System architecture



Arabic dialects, particularly for those of Maghrebi region, is the main reason behind the absence of works dealing with speech recognition for these vernaculars. As aforementioned, we prepared our corpus in 3 dialects in addition to MSA. One part of this corpus, regarding MSA and Moroccan Berber dialect, has already been used in Lounnas et al. (2020). The second part concerning the Algerian Arabic dialect and Moroccan Arabic dialect were recorded by native speakers recently. We summarize in Table 1 the characteristics of this corpus and the recording conditions such as the number of speakers, environment noise and the total number of tokens.

Taking into consideration that the two parts of the corpus have been recorded in conditions different from one speaker to another, we had to re-sample the recorded digits to get a uniform sampling frequency using Praat[2].

Then, we segmented the recorded signals into small fragments. This task is performed using both Praat and Audacity[3].

architectures, one uses acoustic-spectral information, and the other one is based on spectrogram images.

5.1.1 Acoustical-based DI architecture

Our first architecture is four blocks as presented in Fig. 2:

- Input Tier:
  Speech utterances.
- Feature Extraction:
  We extract relevant information based on acoustic and spectral cues.
- Classification Process:
  A set of classifiers based on both machine learning and deep learning are applied to identify the dialects.
- Output Tier:
  The dialect of the speech utterance is identified. The system performance is evaluated using F1 score.

[2] http://www.fon.hum.uva.nl/praat/.

[3] https://www.audacityteam.org.

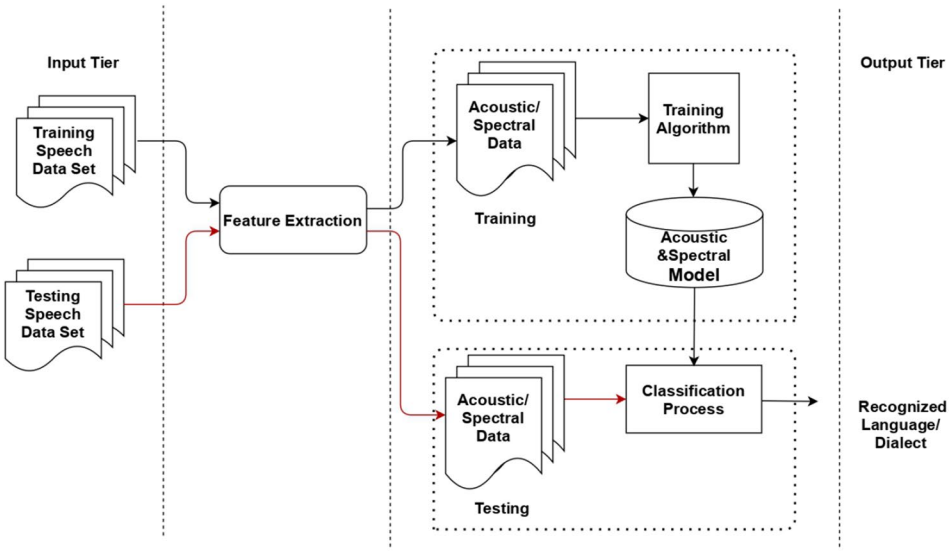


Fig. 2 Acoustic-spectral based DI component

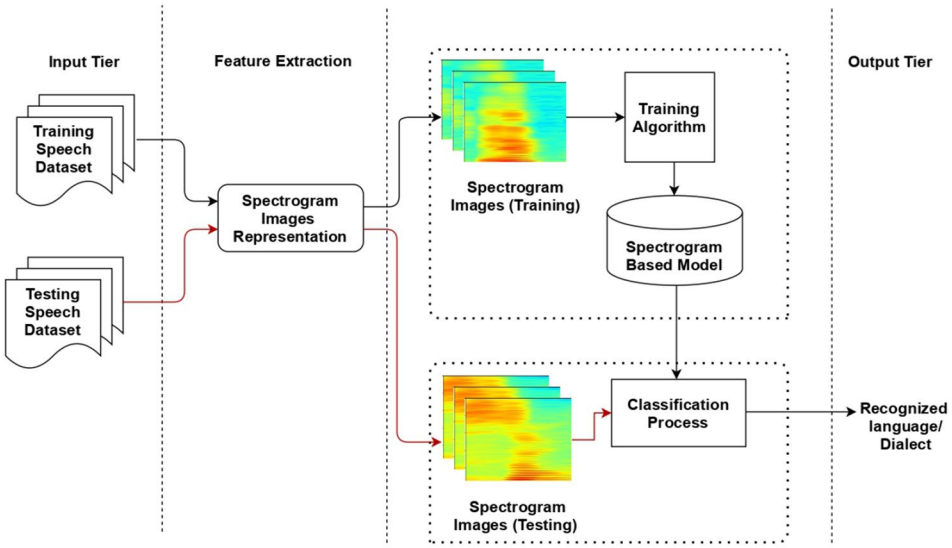


Fig. 3 Spectrogram based DI component

5.1.2 Spectogram-based DI architecture

The input in this architecture is made of a set of spectrogram images of speech signals (Fig. 3).

- Input Tier:
  Speech utterances.
- Spectrogram Representation:
  The spectrogram images are used to train the model.
- Classification Process:
  A set of classifiers based on both machine learning and deep learning are applied to identify the dialects.
- Output Tier:
  The dialect of the speech utterance is identified. The system performance is evaluated using F1 score.

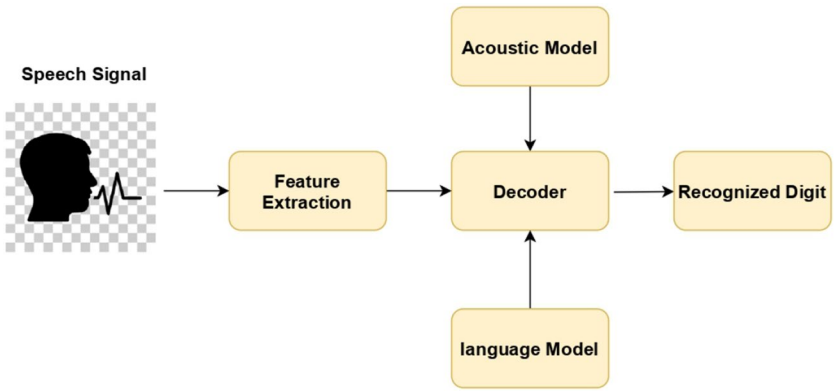
For this purpose, we run several experiments in order to select the classifier that gives the best performance. More details can be found in Sect. 6.

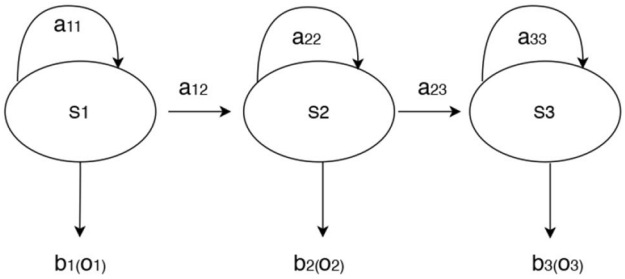
5.2 Automatic speech recognition (ASR)

There are three necessary elements in the ASR system: the acoustic model, the n-gram language model, and the pronunciation dictionary (Fig. 4).

The extracted features are mainly based on the 13-dimensional Mel-Frequency Cepstral Coefficients (MFCC), their delta, and delta-delta vectors. In the decoding phase, the HMM decoder analyzes the features and compares them to the knowledge base. Our ASR system is based on the CMU toolkit (Ezzine et al., 2020; Zealouk et al., 2018) where we used an HMM-GMM approach. Note that each word is

Fig. 4 ASR system





represented as a set of phonemes, and each phoneme is represented by 3-HMM state sequences, one emitting state as an entry and two non-emitting states as an exit that associates HMM units models together in the ASR system. Each emitting state consists of GMMs trained on 39 overall MFCC coefficients. Figure 5 represents our HMM configuration and Table 2 presents the dictionaries related to MSA and the three dialects.

Fig. 5 HMM structure with 3 states

Table 2 The dictionaries used in the training and testing phases

Digits	MBD		MAD	
0	ILEM:	I l E M	SIFER:	SI F E R
1	YEN:	Y E N	WAHED:	W A HH E D
2	SIN:	S I N	JOJ:	J OU J
3	KRAD:	K R A D	THLATA:	TH L A TH A
4	KUZ:	K O Z	RABAA:	R A B AAA
5	SMUS:	S M U S	KHAMSA:	KH A M S A
6	SDES:	S D E SS	STTA:	S TT A
7	SA:	S A	SBAA:	S A B AAA
8	TAM:	T A M	THMANYA:	TH M A N Y A
9	TZA:	T Z A	TSAAOD:	T A S AA OU D
Digits	AAD		MSA	
0	SIFER:	S I F E R	SAFER:	S EY F E R
1	WAHED:	W A HH E D	WAHEDE:	W AA D E
2	ZOUJ:	Z O U J	ETHNAN:	EH TH N AH N
3	TLATHA:	T L A TH A	THLATHA:	TH L AE TH AH
4	REBAA :	R E B AAA	ARBAH:	AA R B AH
5	KHEMSA:	KH A M S A	KHAMSA:	KH A M S AH
6	SETTA:	S E TT A	SETAH:	S EH T AH
7	SEBAA:	S E B AAA	SABAH:	S AA B AH
8	THEMANYA:	TH M A N Y A	THAMANAH:	TH AE M AH N A
9	TESAA:	T E S AAA	TESAH:	T EH S AH

Table 3 Default configuration used for each system

Models	KNN	SVM	EXT	RF	GB
Parameters	Default	RS = 9	NE = 100	NE = 100 RS = 9	RS = 0

RS random state, *NE* number of estimator

6 Experiments and results

In this section, we show the impact of dialect identification on the enhancement of digits spoken recognition for Algerian and Moroccan dialects along with MSA. To that end, we achieved a set of experiments for both dialect identification and speech recognition. To get the best performance for dialect identification, we proposed statistical and deep learning-based approaches.

6.1 Machine learning based dialect identification

6.1.1 Scheme 1: rhythm characteristics, acoustic and spectral features

For the first scheme, we adopted acoustic and spectral features along with rhythm characteristics using the framework[4] based on Librosa[5] (Giannakopoulos, 2015). We present, in the following, the 34 adopted features, namely: MFCC coefficients (13), Energy (1) & Energy of entropy (1), Zero Crossing Rate (1) & Spectral Centroid (1), Spectral Spread (1) & Spectral Entropy (1), Spectral Rolloff(1) & Chroma Vector (12), Spectral Flux (1) & Chroma Deviation (1).

These features are used to train a set of classifiers, namely: k-Nearest Neighbours (KNN), Support Vector Machines (SVM), Extra Trees (EXT), Random Forest (RF), and Gradient Boosting (GB) (Pedregosa et al., 2011). As we aim to select the best features, we used the default configuration of these classifiers (see Table 3). Taking into account the necessity of performing a speaker-independent system, we selected, for each dialect, multiple combinations of speakers to form ten different sets (training and test), in a way we get four speakers representing 65% for training and two speakers representing 35% for the test phase.

Tables 5, 6 and 7 represent performance using the aforementioned ten sets (in Table 4 where S_i denotes speaker number i.) for binary, 3-class, and 4-class classification, respectively.

From Table 5, we noted that regarding 4-class classification using 10 sets (Table 4), GB achieved mostly the best results. We recorded its best performance using the 6th set with an F1-score of 85.89% and an accuracy of 93.03%. Most of the used classifiers achieved their best performances with the 6th set except SVM that yielded its best result using the 2nd set with an F1-score of 78.74% and accuracy of 89.55%.

Table 4 # of different speakers combination

# Set	Training_Speaker's_set	Test_Speaker's_set
01	S1,S2,S3,S4	S5,S6
02	S6,S5,S1,S2	S3,S4
03	S4,S3,S6,S5	S1,S2
04	S4,S3,S6,S2	S1,S5
05	S4,S2,S1,S5	S3,S6
06	S1,S2,S3,S5	S4,S6
07	S1,S3,S4,S6	S2,S5
08	S2,S4,S5,S6	S1,S3
09	S2,S3,S5,S6	S1,S4
10	S1,S3,S5,S6	S2,S4

The 3-class classification gives the best results through the GB classifier with an F1-score of 86.44% and an accuracy of 91.11% (see Table 6).

For binary classification, one can notice from Table 7 that the EXT classifier outperforms the remaining classifiers when dealing with the couples of dialects (AAD-MBD) and (AAD-MAD), with an F1-score of 86.06% and 97.85%, respectively. In addition, it is ranked as the second-best classifier regarding the classification of AAD and MSA with an F1-score of 96.06%. These findings make us to state, intuitively, that the EXT classifier is suitable for inter-class classification (as for instance, Algerian dialect and the Moroccan dialect). For the cases of MBD-MSA, MBD-MAD, and MSA-MAD, the best F1-scores were achieved by SVM (96.78%), GB (93.19%), and KNN (94.99%), respectively. Roughly speaking, the two best overall scores were obtained by the EXT classifier for the AAD-MAD pair, followed by the SVM for MBD-MSA.

6.1.2 Scheme 2: spectrogram

This approach consists of transforming the raw speech into the spectral domain by computing its spectrogram. The set of global characteristics: Hu Moments (Žunić et al., 2010; Sun et al., 2015), Haralick Texture (Sengupta et al., 2019) and Color Histogram (Sergyan, 2008) are retrieved from the spectrograms, already computed and concatenated to form the features vectors. The results presented in Table 8 show, in the case of 4-class classification that the best performance is achieved by GB, with 72.11% (F1) and 86.51% (accuracy). It should be noted that this performance is lower

[4] https://github.com/tyiannak/pyAudioAnalysis.

[5] https://librosa.org/doc/latest/index.html.

Table 5 Obtained results of our dialect identification system based on acoustic and spectral features and rhythm characteristics "4-class classification"

Set\models	KNN	SVM	EXT	RF	GB
1	Acc = 78.93	Acc = 77.5	Acc = 83.03	Acc = 82.85	Acc = 89.19
	F1 = 52.85	F1 = 51.54	F1 = 62.24	F1 = 61.86	F1 = 78.04
2	Acc = 84.28	**Acc = 89.55**	Acc = 88.83	Acc = 87.58	Acc = 92.41
	F1 = 66.79	**F1 = 78.74**	F1 = 76.06	F1 = 76.68	F1 = 84.25
3	Acc = 65.35	Acc = 72.41	Acc = 70.71	Acc = 72.5	Acc = 74.28
	F1 = 29.82	F1 = 39.35	F1 = 37.97	F1 = 41.74	F1 = 44.5
4	Acc = 78.57	Acc = 87.5	Acc = 86.96	Acc = 86.33	Acc = 87.5
	F1 = 57.31	F1 = 75.28	F1 = 73.38	F1 = 72.37	F1 = 74.91
5	Acc = 83.57	Acc = 86.87	Acc = 89.19	Acc = 91.25	Acc = 91.16
	F1 = 66.27	F1 = 72.68	F1 = 76.67	F1 = 81.33	F1 = 81.15
6	**Acc = 89.28**	**Acc = 85.62**	**Acc = 91.33**	**Acc = 91.96**	**Acc = 93.03**
	F1 = 77.88	**F1 = 71.04**	**F1 = 82.47**	**F1 = 83.80**	**F1 = 85.89**
7	Acc = 78.48	Acc = 80.71	Acc = 82.85	Acc = 83.12	Acc = 87.05
	F1 = 55.14	F1 = 61.29	F1 = 65.58	F1 = 65.78	F1 = 73.46
8	Acc = 76.78	Acc = 84.91	Acc = 85.00	Acc = 85.35	Acc = 86.51
	F1 = 53.42	F1 = 69.77	F1 = 70.02	F1 = 70.75	F1 = 72.59
9	Acc = 79.82	Acc = 83.92	Acc = 81.16	Acc = 81.16	Acc = 83.12
	F1 = 55.56	F1 = 66.61	F1 = 60.20	F1 = 60.05	F1 = 63.09
10	Acc = 77.85	Acc = 76.87	Acc = 77.76	Acc = 75.62	Acc = 78.39
	F1 = 55.11	F1 = 54.07	F1 = 53.94	F1 = 47.66	F1 = 55.26

The bold values represent the best results in a given experiment

Table 6 Obtained results of our dialect identification system based on acoustic and spectral features and rhythm characteristics "3-class classification"

Models	KNN	SVM	EXT	RF	GB
Multi-dialects	Acc = 87.14	Acc = 85.39	Acc = 87.93	Acc = 90.47	**Acc = 91.11**
	F1 = 80.12	F1 = 77.35	F1 = 81.18	F1 = 85.32	**F1 = 86.44**

The bold values represent the best results in a given experiment

than that obtained by the former approach (Scheme 1) by around 13.7%. The 3-class classification system, dealing with the three dialects (MBD, AAD, and MAD), has given an F1-score of 87.52% and an accuracy of 91.90% via the RF classifier (Table 9). This can be seen as an improvement of about 1% in comparison to GB performance recorded in Scheme 1.

By analyzing the results displayed in Table 10, which is related to the binary classification case, we note that the best performance is recorded for EXT and RF in the cases of (AAD-MBD, AAD-MAD, and MSA-MAD) and (AAD-MAD, MBD-MSA, MBD-MAD), respectively. Regarding intra-class classification (Moroccan dialects), RF yielded the best performance for (MBD-MAD) pair, in addition to (MBD-MSA).

6.1.3 Scheme 3

In this part, we used Librosa framework[6] (McFee et al., 2015), which includes spectral features and rhythm characteristics. The features used in this framework are composed of 193 components: MFCC coefficients (40), Mel spectrogram (128) & Chroma Vector (12), Spectral contrast (7) & Tonnetz(6).

As shown in Table 11, the best results are performed by EXT with an F1-score and accuracy of 94.46% and 88.13%, respectively. This representation, composed of 193 components, improved F1 score by 3%, compared to Scheme 1 results.

[6] https://github.com/mtobeiyf/audio-classification.

Table 7 Obtained results of the dialect identification system based on acoustic and spectral features and rhythm characteristics "binary classification"

Models	KNN	SVM	EXT	RF	GB
AAD-MBD	Acc = 82.14	Acc = 79.64	**Acc = 86.07**	Acc = 85.35	Acc = 85.35
	F1 = 81.98	F1 = 79.56	**F1 = 86.06**	F1 = 85.34	F1 = 85.32
AAD-MSA	Acc = 91.78	Acc = 77.85	Acc = 88.83	**Acc = 96.07**	Acc = 90.00
	F1 = 91.73	F1 = 77.66	F1 = 96.06	**F1 = 96.42**	F1 = 89.98
AAD-MAD	Acc = 93.92	Acc = 90.35	**Acc = 97.85**	Acc = 97.5	Acc = 96.42
	F1 = 93.91	F1 = 90.29	**F1 = 97.85**	F1 = 97.5	F1 = 96.42
MBD-MSA	Acc = 83.57	**Acc = 96.78**	Acc = 90.71	Acc = 95.35	Acc = 94.28
	F1 = 83.19	**F1 = 96.78**	F1 = 90.64	F1 = 95.35	F1 = 94.27
MBD-MAD	Acc = 83.21	Acc = 90.00	Acc = 86.78	Acc = 88.92	**Acc = 93.21**
	F1 = 82.76	F1 = 89.89	F1 = 86.6	F1 = 88.8	**F1 = 93.19**
MSA-MAD	**Acc = 95.00**	Acc = 86.78	Acc = 92.14	Acc = 92.14	Acc = 89.64
	F1 = 94.99	F1 = 86.72	F1 = 92.13	F1 = 92.12	F1 = 89.64

The bold values represent the best results in a given experiment

Table 8 Results of the dialect identification system based on spectrogram "4-class classification"

Models	KNN	SVM	EXT	RF	GB
Multi-dialects	Acc = 83.57	Acc = 78.92	Acc = 86.64	Acc = 84.10	**Acc = 86.51**
	F1 = 66.4	F1 = 55.72	F1 = 68.27	F1 = 67.01	**F1 = 72.11**

The bold values represent the best results in a given experiment

Table 9 Results of the spectrogram based system "3 class-classification"

Models	KNN	SVM	EXTs	RF	GB
Multi-dialects	Acc = 90.74	Acc = 86.50	Acc = 91.74	**Acc = 91.90**	Acc = 91.42
	F1 = 85.59	F1 = 78.65	F1 = 87.10	**F1 = 87.52**	F1 = 86.87

The bold values represent the best results in a given experiment

Table 10 Results of the spectrogram based system "binary classification"

Models	KNN	SVM	EXT	RF	GB
AAD-MBD	Acc = 83.57	Acc = 80.35	**Acc = 84.64**	Acc = 81.07	Acc = 82.5
	F1 = 83.40	F1 = 80.08	**F1 = 84.51**	F1 = 80.83	F1 = 82.40
AAD-MSA	Acc = 78.57	Acc = 72.5	Acc = 72.14	Acc = 68.57	**Acc = 79.28**
	F1 = 78.35	F1 = 72.24	F1 = 71.97	F1 = 68.25	**F1 = 78.80**
AAD-MAD	Acc = 97.14	Acc = 94.28	**Acc = 99.64**	**Acc = 99.64**	Acc = 99.28
	F1 = 97.14	F1 = 94.27	**F1 = 99.64**	**F1 = 99.64**	F1 = 99.28
MBD-MSA	Acc = 68.92	Acc = 67.85	Acc = 66.07	**Acc = 70.00**	Acc = 65.35
	F1 = 68.10	F1 = 66.88	F1 = 65.84	**F1 = 69.99**	F1 = 65.35
MBD-MAD	Acc = 91.07	Acc = 90.35	Acc = 92.50	**Acc = 93.21**	Acc = 91.78
	F1 = 91.01	F1 = 90.28	F1 = 92.46	**F1 = 93.18**	F1 = 91.74
MSA-MAD	Acc = 81.42	Acc = 82.5	**Acc = 83.21**	Acc = 81.42	Acc = 81.42
	F1 = 81.21	F1 = 81.99	**F1 = 82.80**	F1 = 80.86	F1 = 80.86

The bold values represent the best results in a given experiment

Table 11 Results obtained for the dialect identification system (Scheme 3) "4-class classification"

Models	KNN	SVM	EXT	RF	GB
Multi-dialects	Acc = 90.00 F1 = 80.20	Acc = 92.67 F1 = 85.65	**Acc = 94.46** **F1 = 88.13**	Acc = 93.66 F1 = 86.81	Acc = 93.83 F1 = 87.42

The bold values represent the best results in a given experiment

Table 12 Results obtained for the dialect identification system (Scheme 3) "3-class classification"

Models	KNN	SVM	EXT	RF	GB
Multi-dialects	Acc = 82.22 F1 = 73.62	Acc = 87.14 F1 = 81.13	Acc = 90.95 F1 = 86.42	Acc = 89.84 F1 = 84.76	**Acc = 93.01** **F1 = 89.17**

The bold values represent the best results in a given experiment

Furthermore, the results for 3-class classification are presented in Table 12. The best performance is achieved by GB classifier with F1-score equal to 89.17% and accuracy equal to 93.01%, leading to an improvement of about 2% in comparison to both Schemes 1 and 2.

Features representation used in Scheme 3 has given promising results for binary classification. This can be noticed clearly in Table 13. Let us summarize the results in the following points:

- SVM and EXT performed perfectly regarding four pairs of languages (dialects): AAD-MSA, AAD-MAD, MBD-MSA, and MSA-MAD with F1-score and accuracy of 100%.
- For AAD-MSA pair, almost all the classifiers achieved high performance.
- KNN, SVM, and EXT yielded perfect scores for MSA-MAD pair.
- In the case of AAD-MBD pair, the best performance was achieved by the RF classifier with F1-score and an accuracy equal to 95.71%.
- An overall improvement is performed for all the six pairs of languages/dialects in comparison to Schemes 1 and 2.

6.2 Deep learning based dialect identification

This phase consists of adopting a deep neural network approach (Najafian et al., 2018) using a set of features based on Librosa library with 193 features using a Convolutional Neural Network (CNN) classifier (*Experiment 1*). The parameters we used in the CNN architecture are reported in Table 14. Furthermore, we applied in *Experiment 2*, a transfer learning approach by retraining two Resnet models: Resnet50 and Resnet101, using spectrograms as features.

6.2.1 Experiment 1: Librosa + CNN

The results obtained for 4-class and 3-class classification are presented in Tables 15 and 16, respectively. We noticed an F1-score improvement of around 7%, 23%, and 10% compared to Scheme 3, Scheme 2, and Scheme 1 (baseline). However, performance of 3-class classification decreased in comparison to the three aforementioned schemes by about 22%, 20%, and 19%, respectively.

From the results of the binary classification task presented in Table 17, we note the followings points:

- F1 obtained with CNN architecture reaches is 100 % for the three pairs: AAD-MSA, AAD-MAD, and MSA-MAD.
- For AAD-MBD, CNN outperforms the first and second schemes. However, the most performing technique is Scheme 3.
- For MBD-MSA, the first and third schemes outperform CNN.
- For MBD-MAD pair, the CNN performance is the worst compared to all three schemes.

6.2.2 Experiment 2: spectogram + Resnet + CNN

In this experiment, we tackled the 4-class classification problem by retraining the last layer of Resnet50 and Resnet101 (He et al., 2016). It should be noted that for both Resnet architectures we used the same configuration as explained in Table 18; the only exception is the number of layers per model. Table 19 shows clearly the degraded performance compared to Experiment 1, and that Resnet50 outperforms slightly Resnet101.

Table 13 Results obtained for the dialect identification system (Scheme 3) " binary classification"

Models	KNN	SVM	EXT	RF	GB
AAD-MBD	Acc = 78.21	Acc = 86.42	Acc = 92.14	**Acc = 95.71**	Acc = 83.57
	F1 = 78.19	F1 = 86.40	F1 = 92.11	**F1 = 95.71**	F1 = 83.54
AAD-MSA	**Acc = 100**	**Acc = 100**	**Acc = 100**	**Acc = 100**	Acc = 97.85
	F1 = 100	**F1 = 100**	**F1 = 100**	**F1 = 100**	F1 = 97.85
AAD-MAD	Acc = 97.14	**Acc = 100**	**Acc = 100**	Acc = 91.76	Acc = 99.64
	F1 = 97.14	**F1 = 100**	**F1 = 100**	F1 = 91	F1 = 99.64
MBD-MSA	Acc = 97.85	**Acc = 100**	**Acc = 100**	Acc = 98.21	Acc = 83.57
	F1 = 97.85	**F1 = 100**	**F1 = 100**	F1 = 98.21	F1 = 83.11
MBD-MAD	Acc = 77.50	Acc = 84.64	Acc = 82.5	Acc = 87.50	**Acc = 88.57**
	F1 = 77.48	F1 = 84.53	F1 = 81.94	F1 = 87.30	**F1 = 88.42**
MSA-MAD	**Acc = 100**	**Acc = 100**	**Acc = 100**	Acc = 90.71	Acc = 81.42
	F1 = 100	**F1 = 100**	**F1 = 100**	F1 = 90.63	F1 = 80.76

The bold values represent the best results in a given experiment

Whereas, we notice through Table 20 a little improvement recorded for 3-class classification compared to 4-class classification.

As can be noticed in Table 21, the accuracy achieved for binary classification is ranging from 41.78 to 89.28% (Resnet50) and from 37.85 to 88.92% (Resnet101). The best results have been recorded for the three pairs: AAD-MBD, MBD-MAD, and AAD-MAD. Overall, Resnet101 performance is, in most cases, better than Resent50, except for the pairs: MBD-MSA and MSA-MAD.

6.3 Multilingual ASR baseline system

In order to recognize the ten first digits spoken in MSA, MBD, MAD, and AAD, several experiments, with 3 HMM states and different Gaussian Mixture Models (4, 8 , 16 Gaussians), have been carried out.

On the one hand, we implemented four independent recognition engines for MSA, MBD, MAD, and AAD, respectively.

The best accuracy is obtained by using 3 HMMs and 4 GMMs, as shown in Fig. 6. On the other hand, we designed multilingual ASR baseline engines. Three ASR configurations have been considered to recognize, first, MAD and AAD jointly (mix-sys-1), second, MAD, AAD, and MBD (mix-sys-2), and third, MAD, AAD, MBD, in addition to MSA (mix-sys-3)[7]. Figure 7 presents the recognition rates of the three configurations, with different GMMs values. The best recognition rates are 58.8 %, 56.7 %, and 49.7% for mix-sys-1, mix-sys-2, and mix-sys-3, respectively. The

Table 14 Our best CNN configuration

Models	CNN
The number of layers	8:(4 Conv1D,2 Pooling,1 Dropout,1 Dense)
The size of the input vector	193
The number of the input channels	1
Filter numbers	64–128
Kernel size	3
Pooling size	3
Activation function	ReLu -SoftMax
The probability of dropout	0.5
The size of batches	64
Maximum epochs	10
Loss	categorical_cross_entropy
Optimizer	rmsprop
Number of Neurons	128

Table 15 Performance of the system based on Librosa+CNN for 4-class classification

Models	CNN
Multi-dialects	**Acc = 97.85**
	F1 = 95.75

The bold values represent the best results in a given experiment

Table 16 Performance of the system based on Librosa+CNN for 3-class classification

Models	CNN
Multi-dialects	Acc = 81.90
	F1 = 67.46

[7] mix-sys-1, mix-sys-2, and mix-sys-3: acoustic and language models have been built using a mixture of (MAD and AAD), (MAD, AAD, and MBD), and MAD, AAD, MBD, MSA) corpora, respectively.

Table 17 Performance of the system based on Librosa + CNN for binary classification

Models	AAD-MBD	AAD-MSA	AAD-MAD	MBD-MSA	MBD-MAD	MSA-MAD
CNN	Acc = 87.85	**Acc = 100**	**Acc = 100**	Acc = 90.35	Acc = 86.42	**Acc = 100**
	F1 = 87.67	**F1 = 100**	**F1 = 100**	F1 = 90.26	F1 = 86.38	**F1 = 100**

The bold values represent the best results in a given experiment

Table 18 Our best ResNet configuration

Models	ResNet
The number of Layers	3:(resnet50 ,1 Dropout,1 Dense)
The size of the input vector	$255 \times 153 \times 3$
The number of the input channels	3
Filter numbers	64,256
Kernel size	3×3
Pooling size	7×7
Activation function	ReLu ,Segmoid,elu
Learning rate	0.0001
Momentum	0.9
The size of batches	100
Maximum epochs	10
Loss	categorical_cross_entropy
Optimizer	Nadam
Number of neurons	2048

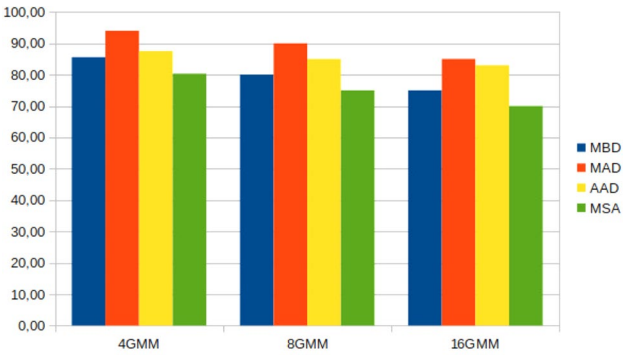


Fig. 6 Speech recognition rates with different GMM

Table 19 Performance of the system based on spectogram + Resnet50/101 + CNN for 4-class classification

Models	CNN
Resnet50	Acc = 75.89
	F1 = 44.74
Resnet101	Acc = 74.73
	F1 = 42.85

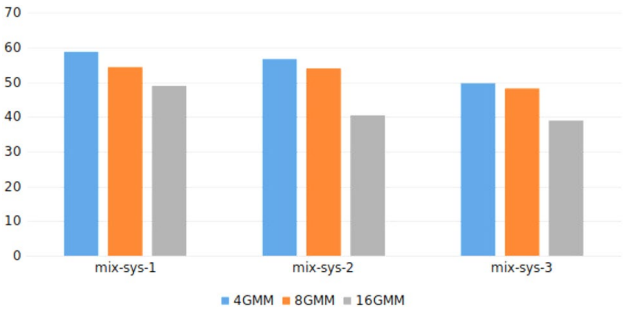


Fig. 7 The accuracy of multilingual ASR baseline system (mix-sys-1, mix-sys-2, and mix-sys-3)

Table 20 Performance of the system based on spectogram + Resnet50/101 + CNN for 3-class classification

Models	CNN
Resnet50	Acc = 77.46
	F1 = 62.09
Resnet101	Acc = 78.25
	F1 = 64.10

Table 21 Performance of the system based on spectogram + Resnet50/101 + CNN for binary classification

Models	AAD-MBD	AAD-MSA	AAD-MAD	MBD-MSA	MBD-MAD	MSA-MAD
Resnet50	Acc = 70.35	Acc = 51.07	Acc = 89.28	Acc = 41.78	Acc = 79.28	Acc = 67.14
	F1 = 67.79	F1 = 35.67	F1 = 89.16	F1 = 39.26	F1 = 78.35	F1 = 63.16
Resnet101	Acc = 75.00	Acc = 51.07	Acc = 88.92	Acc = 37.85	Acc = 79.64	Acc = 58.21
	F1 = 74.84	F1 = 35.67	F1 = 88.79	F1 = 35.07	F1 = 78.76	F1 = 50.08

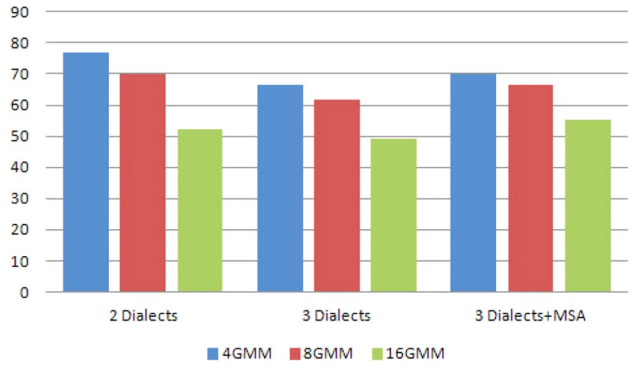


Fig. 8 The accuracy of our proposed multilingual ASR system

best scores are obtained with 4 GMMs. This is probably due to the small number of the used data. The recognition rates dropped dramatically with the increase of the number of dialects to be trained jointly.

To improve ASR systems' accuracy, we integrated the language identification component, which identifies the speaker's language/dialect before the speech recognition process. This will be detailed in the next section.

6.4 The multilingual ASR system

Our proposed system is a combination of Automatic Speech Recognition and Language/Dialect Identification, which is able to switch between the four independent recognizers mentioned in Sect. 6.3 (Fig. 6). It allows selecting the suitable ASR system to recognize the utterance spoken in a particular language/dialect that is identified and provided by the DI module Fig. 1.

As the accuracy of the three ASR engines corresponding to the configurations mix-sys-1, mix-sys-2, and mix-sys-3, was unsatisfactory as shown in Fig. 7, we added the dialect identification component by achieving binary, 3-class, and 4-class classification, according to the number of dialects, considered for each of the three configurations (See Table 22). We notice a significant improvement achieved by our proposed multilingual system using 3 HMMs and 4 GMMs (see Fig. 8 and Table 22) compared to the baseline one.

7 Conclusion

In this paper, we presented a set of experiments for the sake of spoken digits recognition improvement, by adding the language/dialect identification component to standard ASR. We showed that our proposed system is useful for such a task dealing with Maghrebi vernaculars considered as under-resourced languages. We used different approaches for identifying these dialects. In fact, the best performance of 4-class classification (AAD, MAD, MBD, MSA) is achieved using the 3rd scheme, based on Librosa (193 components), to feed the CNN model. The machine learning based classifiers (SVM, EXT, KNN, RF, GB) achieved the best performance, either with Librosa acoustical features or with the spectrogram, when dealing with the three dialects (AAD, MAD, MBD). Overall, dealing with binary or multi-class classification of the dialects, the best scheme is Librosa + CNN, which yielded an accuracy of 100% in some cases, achieved by selecting the appropriate configuration of the CNN model. The second-best performance is achieved by the system based on Librosa with (KNN, SVM, EXT, RF and GB). Using the global features (Hu Moments, Haralick Texture, and Color Histogram) extracted from spectrogram images input, these classifiers outperform Resnet50/101 models that used directly spectrogram images. The latter models are less efficient because of the low number of images.

Our proposed multilingual ASR system has successfully improved the recognition rate of digits spoken in low-resourced dialects from the Maghreb region. In our future research, we will focus on expanding our corpus to cover more dialects.

Table 22 Accuracy of our proposed multilingual ASR system (Fig. 8) compared to the baseline one (Fig. 7) using the best DI system

	3 dialects +MSA	3 dialects	2 dialects
DI system performance	97.85	93.01	100
Proposed multilingual ASR system	69.8	66.37	76.75
Baseline multilingual ASR system	49.7	56.7	58.8

References

Azim, M. A., Hussein, W., & Badr, N. L. (2021). Spoken arabic digits recognition system using convolutional neural network. In *International Conference on Advanced Machine Learning Technologies and Applications* (pp. 164–172). Springer.

Belgacem, M., Antoniadis, G., & Besacier, L. (2010). Automatic identification of Arabic dialects. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA).* Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/719_Paper.pdf.

Bougrine, S., Cherroun, H., & Ziadi, D. (2018). Prosody-based spoken Algerian arabic dialect identification. *Procedia Computer Science, 128,* 9–17.

Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., & Torres-Carrasquillo, P. A. (2006). Support vector machines for speaker and language recognition. *Computer Speech & Language, 20*(2–3), 210–229.

Chittaragi, N. B., Limaye, A., Chandana, N., Annappa, B., & Koolagudi, S. G. (2019). Automatic text-independent kannada dialect identification system. In *Information Systems Design and Intelligent Applications* (pp. 79–87). Springer.

Chittaragi, N. B., Prakash, A., & Koolagudi, S. G. (2018). Dialect identification using spectral and prosodic features on single and ensemble classifiers. *Arabian Journal for Science and Engineering, 43*(8), 4289–4302.

El Ghazi, A., Daoui, C., Idrissi, N., Fakir, M., & Bouikhalene, B. (2011). Speech recognition system based on hidden markov model concerning the moroccan dialect Darija. Global Journal of Computer Science and Technology.

Ezzine, A., Satori, H., Hamidi, M., & Satori, K. (2020). Moroccan dialect speech recognition system based on cmu sphinxtools. In *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)* (pp. 1–5). IEEE.

Giannakopoulos, T. (2015). pyaudioanalysis: An open-source python library for audio signal analysis. *PLoS ONE, 10*(12), e0144610.

Hanani, A., & Naser, R. (2020). Spoken arabic dialect recognition using x-vectors. *Natural Language Engineering, 26,* 691–700.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Kakouros, S., Hiovain, K., Vainio, M., & Šimko, J. (2020). Dialect identification of spoken north s\'ami language varieties using prosodic features. arXiv preprint arXiv:2003.10183.

Lachachi, N. E., & Adla, A. (2016). Two approaches-based l2-SVMs reduced to MEB problems for dialect identification. *International Journal of Computational Vision and Robotics, 6*(1–2), 1–18.

Liu, G. A., & Hansen, J. H. (2011). A systematic strategy for robust automatic dialect identification. In *2011 19th European Signal Processing Conference* (pp. 2138–2141). IEEE.

Lounnas, K., Abbas, M., Teffahi, H., & Lichouri, M. (2019). A language identification system based on voxforge speech corpus. In *International Conference on Advanced Machine Learning Technologies and Applications* (pp. 529–534). Springer.

Lounnas, K., Demri, L., Falek, L., & Teffahi, H. (2018). automatic language identification for berber and arabic languages using prosodic features. In *2018 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM)* (pp. 1–4). IEEE.

Lounnas, K., Satori, H., Teffahi, H., Abbas, M., & Lichouri, M. (2020). Cliasr: a combined automatic speech recognition and language identification system. In *2020 1st International Conference on Innovative Research in Applied Science Engineering and Technology (IRASET)* (pp. 1–5). IEEE.

McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference* (Vol. 8, pp. 18–25). Citeseer.

Mouaz, B., Abderrahim, B. H., & Abdelmajid, E. (2019). Speech recognition of Moroccan dialect using hidden markov models. *Procedia Computer Science, 151,* 985–991.

Najafian, M., Khurana, S., Shan, S., Ali, A., & Glass, J. (2018). Exploiting convolutional neural networks for phonotactic based dialect identification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5174–5178). IEEE.

Nour-Eddine, L., & Abdelkader, A. (2015). Gmm-based maghreb dialect identification system. *JIPS, 11*(1), 22–38.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research, 12,* 2825–2830.

Satori, H., & ElHaoussi, F. (2014). Investigation amazigh speech recognition using CMU tools. *International Journal of Speech Technology, 17*(3), 235–243.

Sengupta, S., Yasmin, G., & Ghosal, A. (2019). Speaker recognition using occurrence pattern of speech signal. In *Recent Trends in Signal and Image Processing* (pp. 207–216). Springer.

Sergyan, S. (2008). Color histogram features based image classification in content-based image retrieval systems. In *2008 6th International Symposium on Applied Machine Intelligence and Informatics* (pp. 221–224). IEEE.

Shon, S., Ali, A., Samih, Y., Mubarak, H., & Glass, J. (2020). Adi17: a fine-grained arabic dialect identification dataset. In *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8244–8248). IEEE.

Sun, Y., Wen, G., & Wang, J. (2015). Weighted spectral features based on local Hu moments for speech emotion recognition. *Biomedical Signal Processing and Control, 18,* 80–90.

Terbeh, N., Maraoui, M., & Zrigui, M. (2018). Arabic dialect identification based on probabilistic-phonetic modeling. *Computación y Sistemas, 22*(3), 863–870.

Touazi, A., & Debyeche, M. (2017). An experimental framework for arabic digits speech recognition in noisy environments. *International Journal of Speech Technology, 20*(2), 205–224.

Wazir, A. S. M. B, & Chuah, J. H. (2019). Spoken arabic digits recognition using deep learning. In *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)* (pp. 339–344). IEEE.

Zealouk, O., Satori, H., Hamidi, M., Laaidi, N., & Satori, K. (2018). Vocal parameters analysis of smoker using amazigh language. *International Journal of Speech Technology, 21*(1), 85–91.

Zerari, N., Abdelhamid, S., Bouzgou, H., & Raymond, C. (2018). Bi-directional recurrent end-to-end neural network classifier for spoken arab digit recognition. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)* (pp. 1–6). IEEE.

Žunić, J., Hirota, K., & Rosin, P. L. (2010). A hu moment invariant as a shape circularity measure. *Pattern Recognition, 43*(1), 47–57.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.