

Arabic dialect classification using an adaptive deep learning model

Nejib Tibi¹, Mohamed Anouar Ben Messaoud^{1,2}

¹Department of Electrical Engineering, Engineering School of Tunis, El Manar University, Tunis, Tunisia

²Department of Physics, Faculty of Sciences of Tunis, El Manar University, Tunis, Tunisia

Article Info

Article history:

Received Jan 10, 2024

Revised Oct 28, 2024

Accepted Nov 19, 2024

Keywords:

Arabic dialect

Deep learning

Deep neural network

Dialect identification

Multi-scale product

Time-frequency domain

ABSTRACT

In daily life, dialect is the most widely used form of communication. Automatically identifying a dialect is a challenging task, particularly when dealing with similar dialects spoken in the same nation. In this study, we developed an automatic dialect identification of feature extraction based on the deep learning model. First, we extract the cepstral features, the fundamental frequency and glottal instances using our multi-scale product analysis (MPA) of the speech signal. These parameter measurements from the MPA of the speech signal are used as features for the designed Hamilton neural network (HNN) classifier. Our classifier considers both the external and the internal dependencies and allows one to code the dependencies by composing the multi-dimensional features as single entities as well as by determining the correlations between the elements by the recurrent operation. Experimental results show that the proposed dialect identification system achieves significant performance gains compared to current HNN-based approaches. The proposed system is rigorously designed to exploit the strong temporal and spectral relationships of speech, and its components operate independently and in parallel to accelerate processing. In addition, the experimental results indicated the robustness of our deep learning model for the identification of Arabic dialect.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Mohamed Anouar Ben Messaoud

Department of Electrical Engineering, Engineering School of Tunis, El Manar University

B.P. 1002 Elbelvédère, Tunis, Tunisia

Email: anouar.benmessaoud@yahoo.fr

1. INTRODUCTION

Dialect identification is the task of recognizing the regional speech variant of a speaker within a given language [1]-[3]. It is an essential research domain in the speech research community because of its significance for automatic speech recognition. Due to phonetic similarities between dialects of a specific language, the challenge of dialect identification has been considered more difficult than that of language identification [4]-[6]. The automatic dialect identification is applied in many domains, such as multi-language translation, classification of document retrieval systems, and automatic speech recognition. An appropriate dialect identification system can help to retrieve and process historical spoken materials to increase their effectiveness [7]. In addition to these, dialect identification can be used in nativity identification, machine translation, text-to-speech synthesis, and media [8]. In the literature, there is a wealth of research on numerous forensic application tasks involving linguistic data such as speaker verification and speaker profiling [9]. These techniques, which examine distinctive voice characteristics and patterns, are crucial for differentiating speakers. In particular, dialect analysis focuses on the subtle differences in

vocabulary, grammar, and pronunciation that are useful in forensic investigations to distinguish speakers from various geographic or linguistic backgrounds.

The phonological, grammatical variations, and lexical language usage that compose dialects are relatively small and subtle. These variances are primarily caused by the distinctive speaking styles used by the group of speakers. The development of one's speaking patterns results from the influence of peripheral elements including culture, locality, social status, and educational background. For automatic dialect identification, numerous studies have shown that utterances contain dialectal information. This is because speakers of various dialects use phonemes differently due to differences in acoustic space. Compared to age, gender, and other relevant factors, dialectal variances are the main cause of speech variability. Dialect identification uses major linguistic variations with specific limits between dialects. It is especially relevant for Arabic dialects. Levantine, North African, Egyptian, Gulf, and modern standard Arabic that is used in informal texts and the Quran are among the most well-known Arabic dialects, each with its own phonetic and linguistic features. Distinguishing between these dialects and even between different languages is possible by leveraging a wide range of information from audio data. This includes lexical, prosodic, phonotactic, acoustic-phonetic, and syntactic elements, which cover the spectrum from low-level sound patterns to higher-level language structures in speech. The smallest units of speech sound in a language are known as phonemes, and the quantity and frequency of phonemes vary among languages. Concatenated phonemes provide phonotactic information, while accentuation, intonation, and rhythm are the three processes that yield prosodic information. Some research has been conducted on the identification of Arabic dialects. Existing research in linguistics examines the grammatical differences between variants of the North African Arabic dialect, the Levantine Arabic dialect, and the Egyptian Arabic dialect. The challenge of identifying between what are probably the most economically significant variations of Arabic gets some interest, especially in shared tasks such as discriminating between related languages. However, some previous techniques for the identification of dialects in Arabic suffer from analytical errors. For example in [10], the authors describe how item names alter their models, diverging from dialect identification due to false correlations in the data used for training, affecting model generalization. Additionally, many studies have concentrated on using speech characteristics at the phonotactic and acoustic levels. On the other hand, few tentatives have attempted to apply prosodic information. Nevertheless, it has been demonstrated that speakers' phonetic and prosodic characteristics, such as glottal instant and fundamental frequency, produce dialect variations.

By addressing this challenge, we want to improve best practices for training dialect identification algorithms. Our research attempts to go deeper into the task of identifying Arabic dialects. To achieve this, we use the performance of deep learning-based models. Additionally, we investigate the features during training and evaluation in order to identify the effect on model performance. In this paper, the multi-scale product analysis (MPA) is used as a pre-treatment for feature extraction. The MPA of speech has a periodic shape with larger singularities denoted by extrema. The multiscale technique has been proposed to improve glottal closure and glottal open instant detection in various dialects. Second, we propose a neural model to identify and classify Arabic dialects. The proposed model has been confirmed to be able to learn the interdependencies and intradependencies between multidimensional input characteristics (glottal closure instant (GCI), glottal opening instant (GOI), open quotient (Oq), F0, mel frequency cepstral coefficient (MFCC), and its derivatives). While previous studies investigated the impact of the acoustic level. Finally, the dialect is classified, and we have demonstrated how our proposed deep learning model can be used to automatically identify the dialects. Our contributions may be summarized as follows:

- We employ the MPA on a non-exhaustive set of useful variables to discern variant Arabic dialects.
- To identify and classify the various Arabic dialects, we applied a causal deep learning-based model for the Arabic dialect identification (ADI) system.
- A deep learning model is applied to learn an efficient representation of the multi-dimensional feature data.

The rest of this paper is organized as follows. Section 2 presents some related works on dialect identification. The classification model is introduced and explained in section 3. Section 4 describes the two datasets used and the results. Section 5 provides conclusions.

2. RELATED WORK

As earlier discussed, dialects can usually be determined using acoustic-phonetic, phonotactic, and prosody levels alone or in combination. To the best of our knowledge, the pioneering work started in the mid-1990s. In the last decade, artificial intelligence, machine learning, and the remarkable ability of the deep learning field have been capable of handling the characterization of speakers' accent. For the identification of accents, various methods have been proposed. For example [11] addressed the challenge of training automatic speech recognition systems. They used auto-encoders based on convolutional neural networks

(CNNs) to train and change the spectrogram. The accuracy of voice recognition has allegedly improved substantially, according to the authors' simulation data. In the work of [12], a Gaussian mixture model was used to classify the dialect of the speech. The authors examined Mandarin corpus data obtained by Asia Microsoft Research. The data included more than 500 speakers identified as male or female, with regional accents from Beijing, Shanghai, and Guangdong, as well as slow, normal, or fast speaking rates. Dialect identification error is achieving an accuracy of 80%. In the work of [13], the authors combined deep neural network (DNN) and recurrent neural network (RNN) to identify the dialects. A combination of short-term and long-term training is applied. Their study revealed that the combination of neural networks and support vector machine (SVM) performed better than using neural networks or SVM alone.

For Arabic dialects, numerous speech datasets have been obtained. Gelly *et al.* [14], used YouTube videos in three different dialects to generate a collection of Arabic speech. They suggested an approach that uses audio-cepstral characteristics as inputs to the universal background GMM. The 2015 challenges for language recognition evaluation presented a dataset of twenty dialects. Ali *et al.* [15], proposed combining DNN on phonotactic characteristics, RNN on acoustic characteristics, and programmable logic device (PLD) analysis on i-vector characteristics to identify the various dialects in the LRE 2015 dataset. The combination of the three models has achieved a score of 0.075. Shon *et al.* [16], have used the ADI dataset. For dialect classification, they applied SVM on both i-vectors and bigram-word and reached an accuracy of 57.20%. In recent research, numerous models have used the ADI dataset to test various dialect classification approaches. Khurana *et al.* [17], the authors achieved a classification accuracy of 73% applying the CNN model, as well as Mel frequency characteristics, before combining the CNN with SVM. Shon *et al.* [18], have also applied the CNN model for MFCC and spectrograms as features. The acoustic convolutional model is combined with the convolutional model as well as the PLD analysis of the i-vectors.

3. METHOD

The proposed approach is detailed in Figure 1. The features include the GCI, GOI, the F0, the MFCC-MPA, and its first, second, and third- order derivatives. The extracted features are applied as input to the proposed Hamilton neural network (HNN) model to make a decision about the dialect of the input speech signal. The HNN model is used to train the model, which is applied for classification.

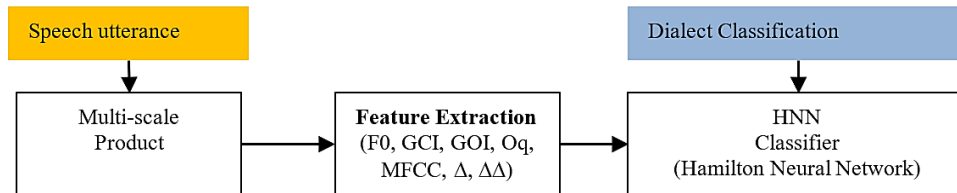


Figure 1. Proposed HNN model

3.1. Multiscale product and wavelet transform

The aim of our application is to determine the features such as GCI, GOI, Oq, and F0 by using the multi-scale product (MP) of the wind turbine (WT). The WT is highly useful for speech analysis tasks such as GCI detection, fundamental frequency determination, speech recognition and synthesis, and so on. The wavelet transform allows for the analysis of speech signals at detailed scales that correlate to the frequency spectrum of speech. In reality, continuous wavelet transform generates maxima. These maxima exist at the mentioned singularity.

To select the appropriate wavelet, we examine two key aspects. However, a one-scale examination does not provide adequate precision. To address this issue, various publications have developed decision algorithms based on various scales [19]. MP was originally used for image processing. The method involves multiplying the WT coefficients at certain scales. Edge features appear at every sub-band in the wavelet domain, and noise diminishes fast as the scales increase. The multi-scale requires multiplying the speech frame $f(n)$ at 3 dyadic scales using the WT coefficients at scale 2^j :

$$p(n) = \prod_{j=j_0}^{j=j_1} w_{2^j} f(n) \quad (1)$$

In this study, the quadratic spline function was applied as a wavelet. Wavelet decompositions are performed at three scales ($s_1, s_2,$ and s_3). Figure 2 illustrates the MP technique.

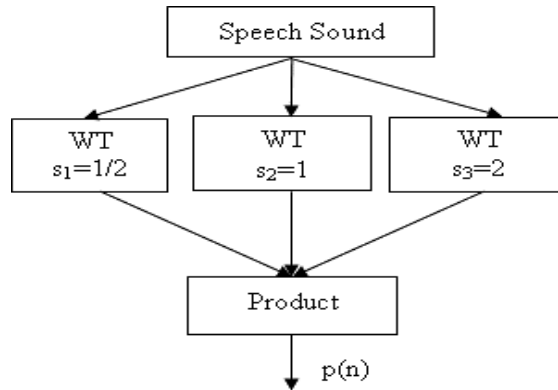


Figure 2. Multi-scale product technique

3.2. Features based on multiscale product

Scale multiplication has been shown to yield better results than other scales, especially in terms of localization performance. We utilize the MP in this work because it produces a derived signal with a simple structure and a shape comparable to the electroglotograph signal. Our technique of determining glottal source signals (GCI, GOI, and Oq), and the F0 is based on further development of our method. The MPA technique is defined as the product of the WT coefficients. We apply a quadratic spline wavelet function that is the derivative of a cubic spline function, and select three dyadic adjacent scales ($s_1 = 2^{-1} = \frac{1}{2}, s_2 = 2^0 = 1$ and $s_2 = 2^1 = 2$) with the levels $j = -1, 0$ and 1 to detect the small peaks. This MP analysis $p(n)$ offers more accurate singularity localization, as shown in Figure 3.

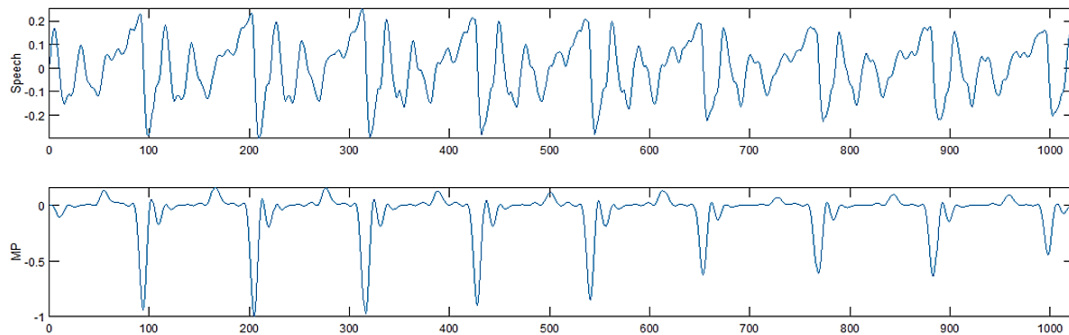


Figure 3. Speech signal followed by the multi-scale product technique

We apply the MP to detect the pitch. Motivated by the efficiency of the MPA in improving the edge detection, we applied our MPA technique to detect the F0, the cepstral feature parameters (MFCC), and its first, second, and third-order derivatives. Thus, we can see that the MP can cancel the additional noise peaks present at its derivative and, consequently, makes a better detection of the fundamental frequency.

3.3. Proposed classification algorithm

For this proposed classification algorithm, our aim is to create an adequate model in order to encode local connections within the characteristics (F0, MFCC-MPA+Δ, MFCC-MPA+ΔΔ, and MFCC-MPA+ΔΔΔ) to avoid the over-fitting and to estimate the minimum number of parameters. To process the entities with more than two related features and to capture these internal latent relations, we apply the Hamilton product system. In this work, the dimensions are divided into 4 parts, or quaternions, which are described in (2):

$$Q = r_1 + x_i + y_j + z_k \quad (2)$$

In (2), each real number has a quaternion unit basis; r_1 is the real part and $x_i + y_j + z_k$ is the imaginary part. The system allows the quaternion-weight elements to be shared by multiple input items. We obtain a quaternion weight:

$$w = w_r + w_x + w_y + w_z \quad (3)$$

We apply the Hamilton product to adapt the hyper-complex quantity of parameters to RNNs in the framework of dialect of speech. Therefore, the neural parameters are applied four times as often as in the standard system. The HNN model makes it possible to deal with the same signal dimension but with four times fewer neural parameters, so we need only four degrees instead of sixteen for a standard neural network. In this work, we propose a modified RNN (Ham-RNN). The forward equation is described as (4):

$$h_t = \alpha(W_{hh} \otimes h_{t-1} + W_{hx} \otimes x_t + b_h) \quad (4)$$

with x_t being the input vector at time t , h_t being the hidden state, α being the activation function, W_{hx} and W_{hh} are, respectively, the input and hidden states.

The complex split activation function α is computed by (5):

$$\alpha(Q) = f(r) + f(x)i + f(y)j + f(z)k \quad (5)$$

with f being a simple split activation function.

Also, we obtain the output vector p_t which is described as (6):

$$p_t = \beta(W_{hy} \otimes h_t) \quad (6)$$

where β is the split activation and W_{hy} is the state weight matrix. First, we construct a non-linear mapping function to determine the regression relationship between speech frames. An absolute error loss function is applied to train the network model. It is described by the following function.

$$l(\theta) = \frac{1}{M} \sum_{l=1}^M |f_{\theta}(Y_l) - T_l| \quad (7)$$

where M is the size of the network training mini-batch size, Y_l is the training feature, and T_l is the speech. Logarithmic frequency density was used as a training feature. It is described by (8):

$$A_{l,k} = \log \left(\frac{S_{l,k}}{x_{l,k}} + 1 \right) \quad (8)$$

In order to stabilize the process of training, we define an amplitude masking for the training target:

$$H_l = \Omega_l \circ \left((1 - Z_l) \circ H_{l-1} + Z_l \circ \tilde{H}_l \right) + (\tilde{F}_l - \Omega_l) \circ H_{l-1} + (\tilde{I}_l - \Omega_l) \circ \tilde{H}_l \quad (9)$$

The raw speech is first separated every 10 ms, with a 25 ms window. The GCI, GOI, Oq, F0, and 40-dimensional log MFCCs with first, second, and third- order derivatives are then recovered using the pytorch-kaldi2 toolkit [20]. An acoustic quaternion $Q(f, t)$ is related to a time frame t and a frequency f as (10):

$$Q(f, t) = P(f, t) + \frac{dP(f,t)}{dt} i + \frac{d'P(f,t)}{dt} j + \frac{d''P(f,t)}{dt} k \quad (10)$$

$Q(f, t)$ describes a number of views of frequency f at time t , including the fundamental frequency, the two glottal instances of the energy at frequency f , and the three derivatives. Quaternions are employed to learn the spatial connections that are present between the three mentioned views that represent the same frequency. Thus, the length of the quadratic input vector is $164/4 = 41$.

4. RESULTS AND DISCUSSION

In this section, we detailed the obtained results and a comparison with other state-of-art methods in dialect classification.

4.1. Datasets

When implementing a system to identify dialects, it is critical to employ training and testing corpora from similar acoustic contexts. For eight Arabic dialects (Tunisian Arabic, Algerian Arabic, Morocco Arabic, Libyan Arabic, Egyptian Arabic, Gulf Arabic, Iraqi Arabic, and Levantine Arabic), we were able to get corpora from the linguistic data consortium (LDC) dataset and the ADI datasets. For the LDC dataset, the recordings consist of spontaneous phone calls between people who speak the dialect as their first language and family members, strangers, and occasionally other people on predetermined topics. For the ADI dataset, it was produced from the Al-Jazeera shows. The statistics for the training and test data are summarized in Table 1.

Table 1. Arabic dialect corpora for training and test for two datasets

Dialect	Train		Test	
	Speakers	Hours	Speakers	Hours
Tunisian	348	11.36	63	2.32
Algerian	567	15.57	102	5.47
Morocco	230	8.49	45	2.44
Libyan	187	8.45	31	1.45
Egyptian	289	11.43	49	1.56
Gulf	356	13.07	63	3.06
Iraqi	371	12.28	75	3.21
Levantine	452	16.31	108	5.32

4.2. Hyperparameters

For training details, the proposed approach has a five-layer structure, with a number of filters equal to 64 filters and a size of filters equal to nine. We employ truncated back-propagation to train the models to increase network training efficiency. Adam optimizer is used to train all networks, with a mini-batch size of 512 and a learning rate of 0.02. All networks are trained with 50 epochs for the dataset, and the final networks are maintained for evaluation. For the test set, two thousand speeches are selected. The speakers are different from those chosen for the training set.

4.3. Performance comparison of networks

Table 2 shows the accuracy, precision, recall, and F1 score of the dialect classification by simple recurrent unit model (RNN), CNN, and proposed approach (HNN), respectively, on the test dataset. We can remark that the three network models can effectively classify the dialect. In addition, the results show that the proposed CNN model outperforms the RNN model in terms of accuracy, revealing that CNN is better suited to dialect identification than RNN due to its deep embedding of the CNN into the RNN. Furthermore, the comparison between the CNN and HNN indicates that the HNN model outperforms CNNR, and achieves higher accuracy, showing the validity of the mechanism of classification by the category used to update the hidden layer characteristics.

Table 2. Performance measures of our approach

Approaches	Accuracy	Precision	Recall	F1 score
RNN	76.41	76.58	76.32	80.56
CNN	81.43	79.67	77.28	81.37
Proposed	85.32	84.65	85.07	88.54

The confusion matrix obtained by our proposed approach is detailed in Table 3. It shows that the Egyptian dialect has a greater accuracy of 89.38%. The reason is that the Arabic dialect spoken in this region (Egypt) frequently has a distinctive speech pattern. As they are neighboring areas of Egyptian dialects, the Levantine dialect exhibits a misclassification rate of 1.31%. Furthermore, it is clear that both Morocco and Algeria are confused with each other. Because these two languages are spoken in adjacent regions of North Africa. The employment of comparable intonation and pitch patterns may be the cause of this. The Iraqi dialect is also misclassified with the Libyan dialect and other Arabic dialects. Iraqi has the lowest classification performance compared to other dialects. This is due to the fact that Iraqi dialects have lower

pitch patterns than other Arabic dialects. Since both testing and validation losses converge at the same time, Figure 4 indicates a good learning rate.

Table 3. Confusion matrix obtained with the proposed approach in ADI dataset

	Tunisian	Algerian	Morocco	Libyan	Egyptian	Gulf	Iraqi	Levantine
Tunisian	89.36	1.25	1.67	0.32	1.73	0.57	1.43	1.02
Algerian	0.23	89.13	0.83	0.73	1.32	0.43	1.76	0.93
Morocco	0.53	0.43	86.43	0.42	1.56	0.78	1.90	1.03
Libyan	0.18	1.32	1.57	88.51	0.93	0.95	1.54	0.94
Egyptian	0.46	1.23	1.33	0.39	89.38	0.82	1.62	1.31
Gulf	0.56	1.63	1.46	1.74	0.14	85.51	0.91	1.27
Iraqi	1.29	2.01	1.91	1.45	0.41	1.43	89.27	1.48
Levantine	0.94	1.98	1.52	1.25	0.56	1.23	1.52	90.42

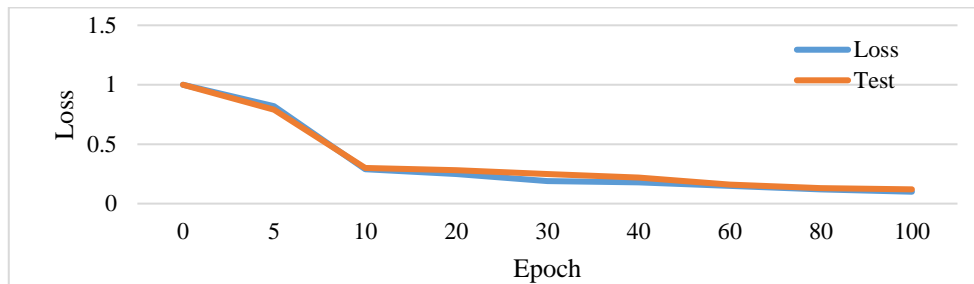


Figure 4. Loss function of the proposed approach

Table 4 shows the comparison results of the accuracy, precision, recall, and F1 scores for the proposed approach and recent deep learning models for the identification of dialects:

- AlexNet [21] is a deep CNN model. Eight layers make up AlexNet, comprising five convolutional layers and three fully linked layers. Dropout regularization and the use of rectified linear units (ReLU) as activation functions were introduced, which helped the network perform better.
- VGGNet19 [22] is recognized for its uniform architecture and simplicity. It used multiple convolutional layers with 3×3 kernels and a max-pooling layer with a 2×2 window, which are present in VGGNet.
- ResNet100 [23] is a DNN architecture created to overcome the problem of vanishing gradients in very deep networks. In ResNet, gradients can flow more easily during training to residual blocks that contain skip connections.
- VFNet [24] was developed to increase the effectiveness and precision of object detection, particularly when detecting objects at various scales. With the aim of improving feature extraction, VFNet makes use of a feature pyramid network (FPN).
- Enhanced support vector machine (ESVM) [25] used a combination of short-term spectrum with chroma features. To capture the variations in timbre, rhythmic, and intonation patterns that are common among dialects, the features aim to collect spectral information.

Table 4. Comparison of dialect classifier approaches

Approaches	Accuracy	Precision	Recall	F1 score
ALI [16]	79.33	83.18	83.74	84.31
KHO [17]	86.22	82.32	84.45	81.89
SHO_V2 [18]	85.51	87.51	86.28	82.23
AlexNet [21]	73.53	72.36	70.27	71.39
VGGNet [22]	74.62	74.25	74.51	72.20
ResNet100 [23]	78.21	77.36	80.68	77.37
VFNet [24]	85.63	81.41	82.21	80.37
ESVM [25]	89.48	88.34	89.56	86.51
Proposed	91.34	90.72	88.27	87.44

It can be shown that the proposed approach outperforms all other approaches in terms of accuracy, precision, recall, and F1 score. This demonstrates that the proposed approach seems better for assigning

dialect identification and classification than the compared models. However, ESVM outperforms the proposed approach by 1.29 for recall. The use of a non-causal model is the main advantage of ESVM over the proposed approach. Non-causal techniques will necessarily have a delay and lack responsiveness in real-time applications. However, the proposed approach uses only current and prior frames.

As can be shown, the proposed approach indicates the best accuracy of 91.34% on the two datasets. This is a 5.83% improvement over the proposed approach described in [18]. The models in [16], [17] also use considerably larger CNN structures and claim final accuracy of 79.33% and 86.22%, respectively.

5. CONCLUSION

In this work, we have proposed an approach that identifies Arabic dialects from a speech. We have designed RNNs to increase the performance of HNN-based dialect identification and classification. The prosodic and acoustic parameters such as GCI, GOI, Oq, F0, MFCC, and its three derivatives have been based on the MPA technique of signals and were applied to extract the features. Then, the designed HNN model was applied to classify the dialect and to give a compact representation for multi-dimensional input features with an efficient learning of feature inter-relations through the Hamilton product. As far as we know, this is the first attempt based on the combination of prosodic and acoustic information in the context of Arabic dialect identification. In addition to this, obtaining dialect information from the target language could enhance the performance of the automatic dialect identification system. More speaker variations can be included by expanding the size of dataset. Other multi-view characteristics that help reduce ambiguity in phoneme representation in the quaternion area will be developed in future work.




REFERENCES

- [1] F. Biadisy, J. Hirschberg, and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling," *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Athens, Greece, 2009, pp. 53-61.
- [2] B. Ma, D. Zhu, and R. Tong, "Chinese dialect identification using tone features based on pitch flux," *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 2006*, pp. I-I, doi: 10.1109/ICASSP.2006.1660199.
- [3] M. J. Harris, S. Th. Gries, and V. G. Migio, "Prosody and its application to forensic linguistics," *Linguistic Evidence in Security Law and Intelligence Journal*, vol. 2, no. 2, pp. 11–29, 2014, doi: 10.5195/lesli.2014.12.
- [4] S. Gray and J. H. L. Hansen, "An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system," *IEEE Workshop on Automatic Speech Recognition and Understanding*, Cancun, Mexico, 2005, pp. 35-40, doi: 10.1109/ASRU.2005.1566480.
- [5] R. Huang, J. H. L. Hansen, and P. Angkititrakul, "Dialect/Accent classification using unrestricted audio," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 453-464, Feb. 2007, doi: 10.1109/TASL.2006.881695.
- [6] G. A. Liu, and J. H. L. Hansen, "A systematic strategy for robust automatic dialect identification," *2011 19th European Signal Processing Conference*, Barcelona, Spain, 2011, pp. 2138-2141.
- [7] M. Abdul-Mageed, A. Elmadany, C. Zhang, El M. B. Nagoudi, H. Bouamor, and N. Habash, "NADI 2023: The Fourth Nuanced Arabic Dialect Identification Shared Task," *Proceedings of ArabicNLP 2023*, Singapore, 2023, pp. 600–613, doi: 10.18653/v1/2023.arabicnlp-1.62.
- [8] A. Faria, "Accent classification for speech recognition," *International Workshop on Machine Learning for Multimodal Interaction*, 2005, pp. 285–293, doi: 10.1007/11677482_25.
- [9] G. Brown, "Y-ACCDIST: An Automatic Accent Recognition System for Forensic Applications," M.S. thesis, Language and Linguistic Science (York), University of York, 2014.
- [10] N. Aepli et al., "Findings of the VarDial Evaluation Campaign 2023," *arXiv*, doi: 10.48550/arXiv.2305.20080, May. 2023.
- [11] K. Radzikowski, L. Wang, O. Yoshie, and R. Nowak, "Accent modification for speech recognition of non-native speakers using neural style transfer," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, p. 11, Feb. 2021, doi: 10.1186/s13636-021-00199-3.
- [12] T. Chen, C. Huang, E. Chang, and J. Wang, "On the use of Gaussian mixture model for speaker variability analysis," in *7th International Conference on Spoken Language Processing (ICSLP 2002)*, ISCA, Sep. 2002, pp. 1249–1252, doi: 10.21437/ICSLP.2002-385.
- [13] R. R. Ziedan, M. N. Micheal, A. K. Alsammak, M. F. M. Mursi, and A. S. Elmaghraby, "Improved dialect recognition for colloquial Arabic speakers," in *2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Dec. 2016, pp. 16–21, doi: 10.1109/ISSPIT.2016.7886002.
- [14] G. Gelly, J.-L. Gauvain, L. Lamel, A. Laurent, V. B. Le, and A. Messaoudi, "Language Recognition for Dialects and Closely Related Languages," in *The Speaker and Language Recognition Workshop (Odyssey 2016)*, ISCA, Jun. 2016, pp. 124–131, doi: 10.21437/Odyssey.2016-18.
- [15] A. Ali, S. Vogel, and S. Renals, "Speech recognition challenge in the wild: Arabic MGB-3," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2017, pp. 316–322, doi: 10.1109/ASRU.2017.8268952.
- [16] S. Shon, A. Ali, and J. Glass, "MIT-QCRI Arabic dialect identification system for the 2017 multi-genre broadcast challenge," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2017, pp. 374–380, doi: 10.1109/ASRU.2017.8268960.
- [17] S. Khurana, M. Najafian, A. Ali, T. A. Hanai, Y. Belinkov, and J. Glass, "QMDIS: QCRI-MIT Advanced Dialect Identification System," in *Interspeech 2017*, ISCA, Aug. 2017, pp. 2591–2595, doi: 10.21437/Interspeech.2017-1391.
- [18] S. Shon, A. Ali, and J. Glass, "Convolutional Neural Networks and Language Embeddings for End-to-End Dialect Recognition," *arXiv*, Apr. 2018, doi: 10.48550/arXiv.1803.04567.




- [19] M. A. B. Messaoud and A. Bouzid, "Pitch estimation of speech and music sound based on multi-scale product with auditory feature extraction," *International Journal of Speech Technology*, vol. 19, no. 1, pp. 65–73, Mar. 2016, doi: 10.1007/s10772-015-9325-1.
- [20] M. Ravanelli, T. Parcollet, and Y. Bengio, "The Pytorch-kaldi Speech Recognition Toolkit," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6465–6469, doi: 10.1109/ICASSP.2019.8683713.
- [21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in neural information processing systems*, Inc., vol. 25, 2012.
- [22] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv*, Apr. 2015, doi: 10.48550/arXiv.1409.1556.
- [23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [24] A. Ahmed, P. Tangri, A. Panda, D. Ramani, and S. Karmakar, "VFNet: A Convolutional Architecture for Accent Classification," in *2019 IEEE 16th India Council International Conference (INDICON)*, Dec. 2019, pp. 1–4, doi: 10.1109/INDICON47234.2019.9030363.
- [25] N. B. Chittaragi and S. G. Koolagudi, "Dialect Identification using Chroma-Spectral Shape Features with Ensemble Technique," *Computer Speech & Language*, vol. 70, pp. 1–14, Nov. 2021, doi: 10.1016/j.csl.2021.101230.

BIOGRAPHIES OF AUTHORS



Nejb Tibi    is graduated in 2010 from ENISO with a degree in Electrical Engineering. In 2019, he graduated from ENIT Tunis, Tunisia, with a Master's degree in Signal System and Data (SSD). He is currently employed at El Manar University in Tunisia as a Research Professor in the Laboratory of LSITI-Department of Electrical Engineering, Engineering School of Tunis (TUNIS). Speech processing is one of his research areas. He can be contacted at email: tibinejb@yahoo.fr.



Mohamed Anouar Ben Messaoud    has been a member of the faculty since 2011 and is currently an associate professor in the Department of Engineering School of Tunis, Faculty of Sciences of Tunis, El Manar University, Tunisia. He has authored or co-authored more than 50 research publications in the fields of speech processing, image analysis, and biomedical engineering, with a focus on neuro-signal processing. He can be contacted at email: anouar.benmessaoud@fst.utm.tn or anouar.benmessaoud@yahoo.fr.