IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 February 2024, accepted 4 March 2024, date of publication 7 March 2024, date of current version 14 March 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3374768*

RESEARCH ARTICLE

# Recognition of Arabic Accents From English Spoken Speech Using Deep Learning Approach

**MANSOOR HABBASH[1], SAMI MNASRI[1,2], MANSOOR ALGHAMDI[1], MALEK ALRASHIDI[1], AHMAD S. TARAWNEH[3], ABDULLAH GUMAIR[4], AND AHMAD B. HASSANAT[3], (Senior Member, IEEE)**

[1]Applied College, University of Tabuk, Tabuk 47512, Saudi Arabia
[2]CNRS-IRIT Laboratory, University of Toulouse II, 31058 Toulouse, France
[3]Faculty of Information Technology, Mutah University, Kerak 61710, Jordan
[4]College of Computer Science, University of Tabuk, Tabuk 47512, Saudi Arabia

Corresponding authors: Sami Mnasri (smnasri@ut.edu.sa) and Ahmad B. Hassanat (Hasanat@mutah.edu.jo)

The authors extend their appreciation to the Deanship of Scientific Research at University of Tabuk for funding this work through the research no. S-1442-0198.

**ABSTRACT** Accents, or changes in how different people speak the same word/sentence in the same language, pose substantial communication issues in most spoken languages. This is a well-known fact, but how does the accent of one language affect learning/speaking another? In this paper, we look at how Arab accents influence the English language. To that end, we built a deep machine-learning system for Arabic accent recognition that was learned from an in-house English speech database of four Arabic accents collected from Jordan, Iraq, Saudi Arabia, and Tunisia. The proposed system employs Mel spectrograms of an English-spoken paragraph to train an LSTM neural network to recognize the accent in each sound signal. Although the collected data was extremely difficult to learn due to the presence of both males and females and fluent speakers in each class, the proposed system could recognize speakers with various accents by up to 79%. This answers the study's main question, demonstrating that speakers with an Arabic accent have their way of speaking English, which varies by country. As a result, if trained on appropriate and adequate data, the proposed system can also be used to recognize accents in any language.

**INDEX TERMS** Accent recognition, classification, deep learning, LSTM.

## I. INTRODUCTION

The accent is a major communication challenge for most spoken languages. In its simple definition, the accent of a language is the fact that different people pronounce the same word in the same language differently. In some conversations, even intuition is not enough to identify the correct meaning of words pronounced in different accents. This causes issues with Automatic Speech Recognition (ASR) systems that do not correctly understand words pronounced by accents other than the standard one.

In general, for learners speaking a second language (L2), the semantics of the language is easier to acquire than the grammar [1]. Hence, pronunciation is an essential skill in language learning, as it facilitates communication. On the other hand, pronunciation, according to [2], is the major difficulty encountered by learners of English as a second language. Moreover, by saying that pronunciation is the ''Cinderella of language teaching'', [3] emphasized the importance of pronunciation when learning a language compared to other skills such as vocabulary or grammar.

Richards (1974) introduces the theory of the interlanguage phase which links the target language (TL) of the learner and his mother tongue (NL). This phase causes errors generated by this linguistic transfer. Several factors affect the severity of these errors. According to [4], among others, these factors concern the pedagogical ways of language teachers, learning plans or even textbooks. Subsequently,

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeeb Dey.

many learners manage to acquire the necessary skills in vocabulary and grammar but are unable to conduct direct conversations due to pronunciation deficiencies. However, this interlanguage phase is usually temporary as the learners improve their pronunciation abilities which is, inevitably, characterized by their TL skills.

In this context, Arabic speakers of English often make spelling errors rather than grammar or syntax errors. Regarding the number of people, the Arabic language is the fifth language spoken in the world [5]. However, few research works are interested in the recognition of Arabic accents, especially from the speeches of other languages such as English. This is mainly explained by the lack of public corpora of data and by the phonetic and syntactic complexity of this language [6]. The countries that speak Arabic are more than twenty. As a result, numerous Arabic dialects exist with distinct pronunciations that make recognizing accents more difficult. Moreover, given this diversity of Arabic accents and regions, the Arabic pronunciation of many English words is different. The paper deals with the pronunciation of English by Arab individuals from different countries.

The Arabic alphabet involves three pairs of vowel phonemes and 28 consonants. Each letter in this alphabet represents a phoneme. In addition, a matching between letters and phonemes exists, which implies that words are generally pronounced as they are written, particularly those, which are fully-vowelized.

To better understand the differences between Arabic and English languages, [7] studied the sound system of the Arabic language and revealed this one-to-one correspondence between Arabic phonemes and letters, differently from English. For the English language, 20 vowels and 24 consonants exist but they are summed up in only 26 letters representing the phonemes. Moreover, the matching between letters and phonemes does not exist and different ways of representing each phoneme are possible.

However, another issue that characterizes the pronunciation of the English language, is that several sounds can be associated with the same letter. Indeed, Arabic learners of the English language often have issues, especially in pronunciation, which can remain even after a long period of learning or practicing conversations in English. Other issues are added to the problem of detecting Arabic accents such as interference and direct transfer from the Arabic language into English as a second language.

Several recognition techniques can be used for accent detection. Huge advances were achieved in voice recognition techniques. This advance has benefited from the progress in data science and deep learning. These advances are both on industrial applications and on academic published papers. Despite this, the reliability and rapidity of ASR systems still need to be improved. The preprocessing systems and the deep learning algorithms for Speech recognition should give high accuracy rates in a reasonable time.

The process of ASR systems relies on separating the individual sounds of audio of a speech, then using algorithms for analyzing the obtained sounds and identifying the most suitable words in the concerned language or accent. The most popular methods used for analyzing the signal and visualizing its sound components are the Discrete Fourier Transform (DFT) [8], the short-time-fourier transform (STFT) [9], and the mel spectrogram.

For the prediction algorithm, short-term memory (LSTM) [10] is one of the most Deep Learning (DL) algorithms used for accent detection. LSTM networks are widely used for processing and classifying data having temporal relationships between its components. LSTM differs from feed-forward neural networks in the involvement of feedback connections. Apart from single independent data such as images, this type of recurrent neural network (RNN) can manipulate data sequences such as audio and video. For this, LSTM is adequate for the processing and recognition of accents from audio sequences.

The main research question allowing us to identify the contribution of the proposed system is:

To what extent do Arabic accents influence the English language? To answer this major question, we must first address sub-major issues such as:

What algorithms are used for ASR systems? What are the most efficient methods for training, extracting features, and recognizing accents? How to propose such an efficient system dedicated to identifying the accent of an Arabic speaker of English as a second language?

The goals of this research involve answering the issues indicated above as well as the following:

- Automatic recognition of foreign accents is important for numerous speech systems, such as voice conversion, speaker identification, and speech recognition.
- Training a deep learning model to automatically detect Arabic accents from English audio speech to help sociolinguists and discourse analysts.
- Investigating various dialects of Arabic spoken in various cultural areas, such as North Africa, the Gulf, the Levant, and Iraq. And creating a new Arabic-accented speech dataset for this purpose.

## II. RELATED WORK

In this section, the most relevant research works studying the English accent detection of Arab speakers, are investigated.

In [11], the focus is set on how Arabic advanced learners acquire grammatical capabilities from indirect questions of English. Ten persons and four native speakers of English were involved to assess oral and written grammatical tasks. The results of this paper indicate some findings that are opposed to the initial hypotheses. Indeed, even linguistic properties that do not link discourse and syntax were problematic for L2 advanced learners. Despite these important statements, this study has the disadvantages of using a small number of people and using a single variant of Arabic learners (Omanis).

The study in [12] investigates mixed speech that occurs using the vocabulary of two or more languages. This paper automatically identifies and recognizes the speech of two languages, English and Sudanese Arabic. The authors investigate the effects of the Sudanese Arabic dialect as a mother tongue on the pronunciation of English. This study proposes a generalized framework for ASR in mixed speech mode (ASR-MS). The used automatic recognition framework considers mixed speech as a new hybrid language. One hundred Arabic Sudanese who mixed Arabic and English sentences in their daily life are involved in the study.

The study in [13] investigates the errors made by Arab Saudi students when pronouncing English. The students were divided into two groups: a group of English major students and a second one of Arabic major students. The study concludes that, as expected, Arabic major students made more errors in pronunciation than English major students. Moreover, the study reveals that Arabic students are involved in a direct transfer from the Arabic language in their English speech pronunciation. They use stress shifts that are not known as English stress patterns. Among the drawbacks of this study, are the limited results, and the non-comparison of the performance of the proposed model with other models. The purpose of the study in [14] is to estimate the age of a speaker by a listener and to identify the effect produced by the native language of the speaker (Arabic, Korean, and Mandarin) when speaking English. Both speakers and listeners have English as a second language. Indeed, native Mandarin, Korean, and Arabic listeners try to estimate the age, to the year, of other persons by listening to recordings of native Mandarin, Korean, and Arabic speakers of English. The study reveals that Mandarin (but not Korean, as assumed in the hypothesis of the study) speakers were perceived to be younger than Arabic speakers. Indeed, it is concluded that the estimation of age becomes more inaccurate if the listener and the speaker have two different linguistic backgrounds.

The study in [15] investigates the efficiency of the systems of online English learning dedicated to non-English speakers. 99 Arabic non-English language students were involved in the experiments. The taken measures concern the percentage of students completing all the vocabulary activities and the percentage of the activities achieved by all the students. Three online systems were tested: adaptive, adaptable, and static. The findings of the study indicate that the usability of the three systems is comparable at the individual and collective levels. However, the learning achievement is different: Indeed, the static system gives less achievement compared to adaptable and adaptive systems.

Some studies investigate the problem of accent detection using deep learning paradigms as follows: In [16], the focus is put on how to use neural models for the attenuation of the differences between accents when establishing an L2 English MDD system using end-to-end (E2E) neural models. The aim is to develop accent-sensitive neural modules based on the fine identification of acoustic differences, to endow the resulting MDD model with a better ability to discriminate these differences.

The study in [17] suggests a set of deep neural models to classify the most known English dialects. The used deep classifiers are the time-delay neural network (TDNN), the convolution neural network (CNN), the temporal convolution neural network (TCN), and the TDNN with emphasized channel attention (ECAPA-TDNN. It was shown by experimental results that the best dialect classifier is ECAPA-TDNN. However, the suggested feature should be extended to perform speaker identification.

The following is an overview of recent studies on the identification of English accents in speakers of languages other than Arabic:

In [18], the authors use audio samples to create models for identifying local accents in the Bengali language. RNN, CNN, and Multilayer Perceptron (MLP) are implemented for creating the models. Despite the reduced error rate of the models, an ASR system should be implemented to differentiate the numerous existing Bengali accents.

The study in [19] shows, from the observation of the accent of English-speaking Mandarins, that Mandarin learners can judge their peers severely even if the content of the speech is fully understood. Another experiment is performed based on the identification of stress from a word in a sentence using a set of tests for word recognition. In the latter experiment, several participants rated the L2 accent excerpts more severely than the native speaker excerpts.

Another research in [20] suggests developing an ASR system for the Kazakh language. The latter has no available public speech corpora. Therefore, the contribution of the authors was to collect sufficient vocal data to implement a reliable Kazakh accent identification environment.

The study in [21] investigates the performance of recognition systems in detecting and analyzing the errors of pronunciation of non-native English speakers (L2). The proposed system of mispronunciation detection and diagnosis (MDD) is based on the use of Electromagnetic Articulography (EMA) data. The acoustic characteristics used to evaluate the performance of the introduced MDD system are based on the Electromagnetic Articulation Corpus of Mandarin Accented English (EMA-MAE).

The work of Kethireddy et al. [22], focused on learning filterbanks that are initialized using customized features obtained from raw waveform, which are incorporated into the CNN network for English accent recognition. The experimental results demonstrate strong performance with an accuracy of 81.26%; these results were attained by using their techniques on a common dataset of 8 English accents. Similar works include [17], [23], [24], [25].

The works of [26] and [27] are maybe the most comparable to ours, thus they will be chosen as the baseline comparisons for the proposed work. To identify the native languages of non-native English speakers from various countries in the Arabic region, such as Saudi Arabia, Egypt, and Tunisia, Mnasri and Habbash [26], proposed a hybrid multi-agent reinforcement learning algorithm. This algorithm makes use of cooperative agents and multi-agent communication. Results from the investigation indicate an average accuracy of 61%.

On the other hand, Ali et al. [27], studied both generative and discriminative models, combining these features with a multi-class SVM. And then validated their findings on distinguishing between the five most common dialects of Arabic, namely Egyptian, Gulf, Levantine, North African, and Modern Standard Arabic (MSA), with an accuracy of 59.2%.

Modern Standard Arabic (MSA), as was shown in this review, is a low-resource language. This also applies to the several Arabic accents that have evolved from MSA, which, despite being frequently spoken, are likewise thought of as low-resource languages and have fewer studies and datasets than MSA. Although we found some online datasets for various Arabic accents, they were exceedingly few and unsuitable for deep learning. By producing a brand-new, comparatively sizable dataset for several Arabic accents speaking English, this paper closes the gap. In addition to presenting a new deep-learning approach that uses state-of-the-art tools to identify various Arab accents when speaking English. This study also explores the impact of these various accents on English language learning, similar to the work of [26] while utilizing more modern techniques and a bigger dataset.

## III. METHODOLOGY

The goal of ASR is to comprehend human speech, therefore identifying accents is essential to enhancing speech recognition. Accented speech recognition can benefit from mature ASR techniques. Although the DFT has been around for a while, it does not include temporal information after transformation [28]. To solve this, the STFT was developed [29], which computes DFT for each time frame in which the signal is divided. STFT preserves frequency and temporal information, which is graphically represented as a spectrogram.

Spectrograms (Figure 1) are thought to be a useful method for determining the strength of a signal over time. This type of visualization has been used successfully in a variety of applications, including voice recognition. Spectrograms are commonly used in machine learning by saving them as image representations for each signal in a dataset, such as speech, and then training a CNN on these images [30].

### A. MEL SPECTROGRAM

Another method for visualizing sound signals is the Mel spectrogram. The Mel spectrogram is interesting in that it provides representations of the sound signal after converting the frequencies to the Mel scale. The bands in the Mel spectrogram are evenly spaced, allowing it to effectively emulate the human ear [31]. The sound signal is transformed in this spectrogram using the linear cosine transform of a log power spectrum.

### B. THE PROPOSED METHOD

In this work, we use Mel spectrograms to translate sound sources from our in-house dataset into psychoacoustic representations. These spectrograms are organized as a three-dimensional array S(B,T,F), with B representing the number of sound signals, T denoting the time steps in the Mel spectrogram, and F representing the features (frequencies) at each time step. The Mel spectrograms are used to train an LSTM neural network for accent recognition, as shown in Figure 2.

The accent recognition process begins with initial speech recordings, which are then used to compute STFT and create Mel spectrograms—a stage known as feature extraction. These spectrograms are used as input to train the LSTM model. Following training, the model is applied to additional instances using the same feature extraction method. Furthermore, while Mel spectrograms may be interpreted as images, LSTM is preferable because it can handle sequences of time-related data, even if they are presented as images.

Algorithm 1 shows the Pseudocode to compute the Mel spectrogram from an audio signal using the Librosa library in Python. The input parameters are as follows: *y* represents the audio signal, *sr* is the sampling rate, *n_fft* is the length of the DFT window (default value is 2048), *hop_length* is the number of samples between successive frames (default value is 512 samples), *window* is the window function type (default is 'hann'), *center* specifies whether to center the frames (default is True), *pad_mode* is the padding mode (default is 'constant'), and *power* is the exponent for the magnitude computation (default value is 2.0). The algorithm then computes the STFT of the audio signal, calculates the magnitude spectrogram, and finally computes the Mel spectrogram using the magnitude spectrogram and the provided sampling rate. One can visit the official package website for more details.[1]

Then, The spectrograms for the recordings in the dataset are saved as images and passed to the proposed LSTM model.

### C. EXPERIMENT DESIGN AND DATASET

We use a new accented speech dataset collected for this study due to a lack of accented Arabic speech recognition corpora, particularly those with Arab subjects speaking English. Subjects from four Arab countries, namely Jordan, Iraq, Saudi Arabia, and Tunisia, contributed to this in-house dataset. Each dataset subject was asked to record his or her reading of the script below:

*" Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station"* [32].

We chose this specific/standard script because it is used in many studies on human-accented speech, such as [33] and [34]. A similar Kaggle speech dataset is available online, but the number of Arabic recordings is very small

[1] https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.html

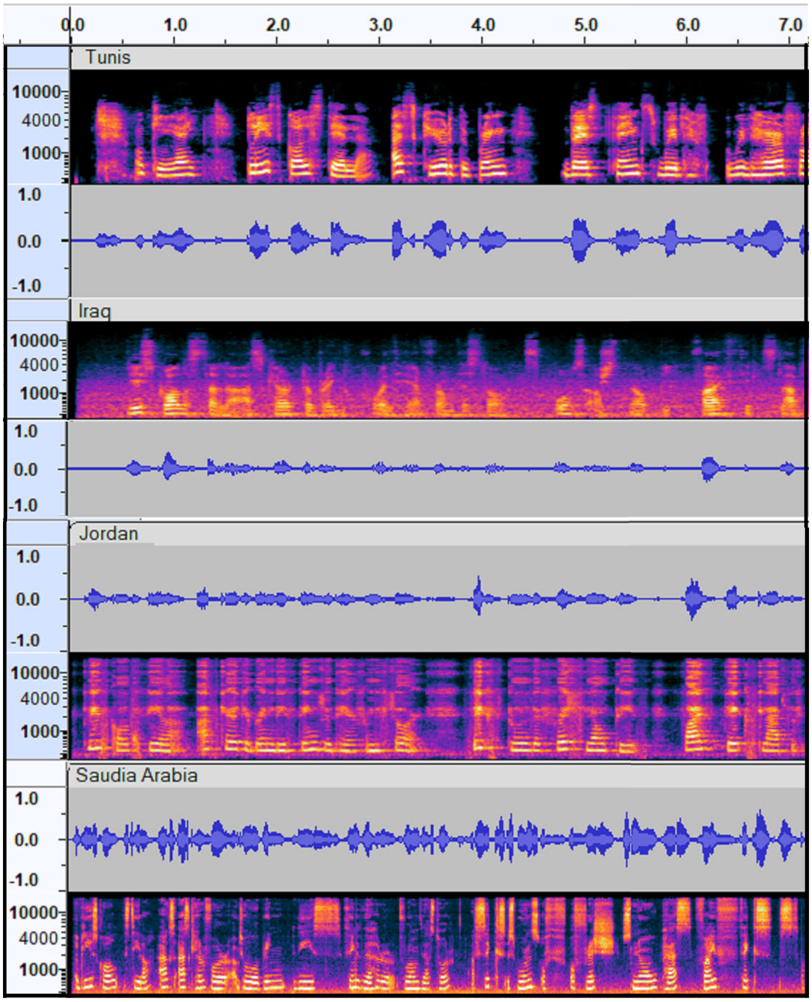


**FIGURE 1.** Spectrogram examples of four speakers from our in-house speech dataset.

**Algorithm 1** Pseudocode for Computing Mel Spectrogram

```
Require: $y$, $sr = 44100$, $n\_fft = 2048$, $hop\_length = 512$, $window$ ='hann', $center = True$, $pad\_mode$ ='constant', $power = 2.0$
Ensure: Mel spectrogram $mel\_spec$
  $stft \leftarrow$ librosa.stft($y, n\_fft, hop\_length, window, center, pad\_mode$)
  $magnitude\_spec \leftarrow |stft|^{power}$
  $mel\_spec \leftarrow$ librosa.feature.melspectrogram($S = magnitude\_spec, sr$)
  return $mel\_spec$
```

(13 Tunisians, 12 Iraqis, 5 Jordanians, and 97 Saudis), making it insufficient for deep learning. Therefore, we had to record our dataset, as listed in Table 1.

The lengths of voice records ranged from 21 to 52 seconds, depending on the reading speed of each subject. Each subject recorded the aforementioned paragraph using their mobile phones, resulting in different file formats. As a result, using Audacity (an open source, cross-platform software for recording and editing sounds) version 3.1.3, we converted all of the voice files into MP3 format, which significantly reduced file sizes and allowed us to unite our dataset to have each record with a sample rate of 44100 Hz and 32 bits per sample since it is suitable for capturing the whole human hearing range (20 Hz to 20 kHz) [35]; using greater increases the data size, while using lower loses information crucial for recognition. Figure 1 shows the spectrogram and waveform samples of four subjects belonging to the four countries mentioned.

The experiments are mainly designed to test the ability of the proposed method to distinguish one accent from another. In all the experiments, we use a hold-out set, with 0.33 for the testing set and 0.67 for the training set, where each

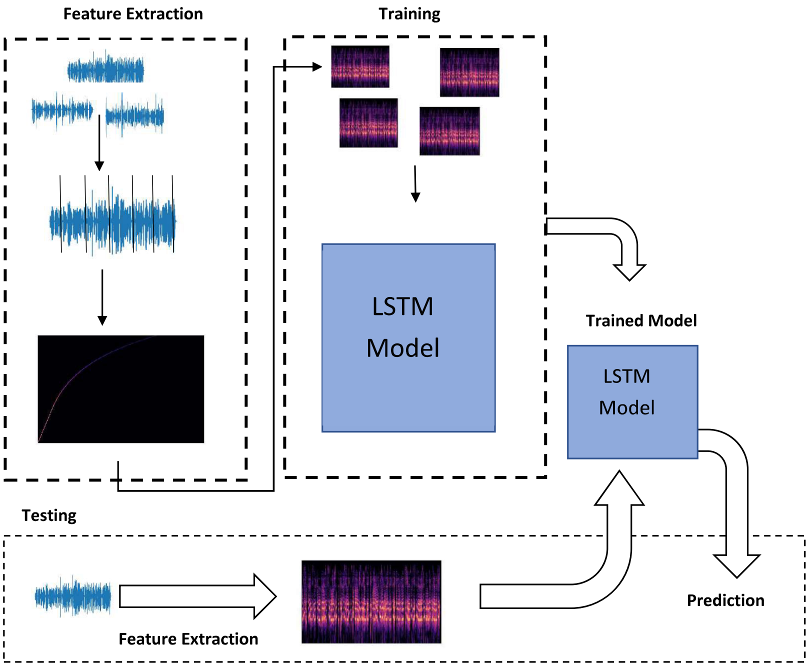


**FIGURE 2.** Diagram of the proposed accent recognition system.

**TABLE 1.** Distribution of our in-house accented speech recordings across countries, and some statistical characteristics.

| ID | Country | No. of records | Female | Male | Fluent | Non-fluent |
|---|---|---|---|---|---|---|
| 1 | Tunisia | 100 | 64 | 36 | 35 | 65 |
| 2 | Jordan | 100 | 37 | 63 | 52 | 48 |
| 3 | Iraq | 91 | 30 | 61 | 30 | 61 |
| 4 | Saudi Arabia | 100 | 41 | 59 | 48 | 52 |
|  | Total | 391 | 172 | 219 | 165 | 226 |

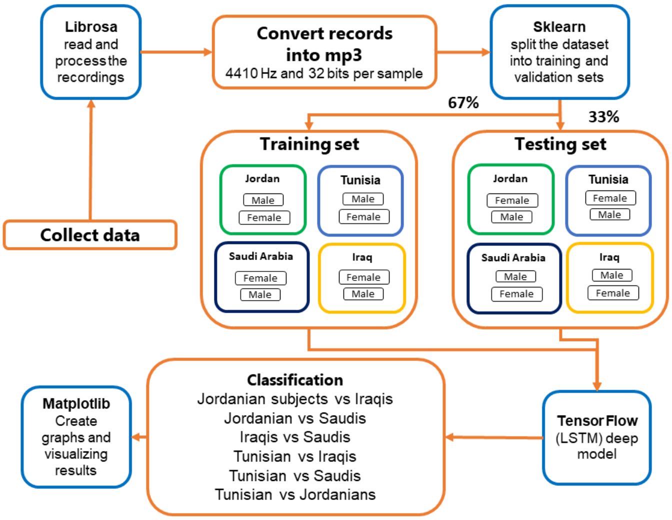


**FIGURE 3.** Flow chart of the experimental setup and processes.

experiment is repeated five times and the average is reported [36], [37], [38], [39].

We run our experiments on Google Colab, which is a Python-based programming environment [40]. Professional packages are used, for example, TensorFlow is used to build LSTM deep model, Librosa is used to read and process the recordings, Sklearn is used to split the dataset into training and validation sets, and Matplotlib is used for graphs and

**TABLE 2.** The architectures of the used deep neural networks. For training, the Adam optimizer is used with 0.001 learning rate and the $n_c classes$ differs based on the experiment.

| Network | Layers |
|---|---|
| LSTM Model | LSTM(128)<br>LSTM(128)<br>LSTM(256)<br>LSTM(256)<br>Dense($n_c lasses$) |
| Bidirectional LSTM Model | Bidirectional(LSTM(128))<br>Bidirectional(LSTM(128))<br>Bidirectional(LSTM(256))<br>Bidirectional(LSTM(256))<br>Dense($n_c lasses$) |
| Conv1D Model | Conv1D(32, kernel_size=3)<br>Conv1D(32, kernel_size=3)<br>Dropout(0.5)<br>MaxPooling1D(pool_size=2)<br>Flatten()<br>Dropout(0.2)<br>Dense(64)<br>Dense($n_c lasses$) |
| MLP Model | Dense(64)<br>Dense(128)<br>Dense(256)<br>Dropout(0.2)<br>Dense(128)<br>Dense(64)<br>Dense($n_c lasses$) |

visualizing the results. Figure 3 illustrates our experimental setup and processes.

## IV. RESULTS AND DISCUSSION

Because the goal of this study is to investigate the impact of different Arabic accents on the English language, we built the proposed accent recognition system to see how much Arabic accents affect the way words and sentences are pronounced in English.

To achieve this purpose, we commenced pilot studies to assess the performance of the proposed system in identifying the accents of speakers from four Arab nations. The proposed method's findings, as well as those of other machine learning algorithms such as MLP, 1-D convolutional neural network (1DCNN), and bi-directional LSTM (BiLSTM), were less than encouraging. Table 2 shows the architecture of these models. The proposed approach achieved the highest accuracy, around 46.6%, as shown in Figure 4. We relate such poor results to challenges caused by variances within a class as well as similarities across the four classes in pronouncing the same sentences and vocabulary. Notably, speakers from Jordan, Iraq, and Saudi Arabia have close Arabic accents, which may manifest in their English-speaking style. The confusion matrix in Figure 4 shows the result of the multi-class classification of the four groups investigated in this study.

**TABLE 3.** The best classification accuracy for each experiment compared to baseline methods.

| Experiment | Best accuracy [%] |
|---|---|
| Jordanians vs Iraqis | 54 |
| Jordanians vs Saudis | 61 |
| Iraqis vs Saudis | 57 |
| Tunisians vs Jordanians | 73 |
| Tunisians vs Iraqis | 79 |
| Tunisian vs Saudis | 70 |
| Eastern vs Western Arabs | 72 |
| [27] | 61 |
| [28] | 59.2 |

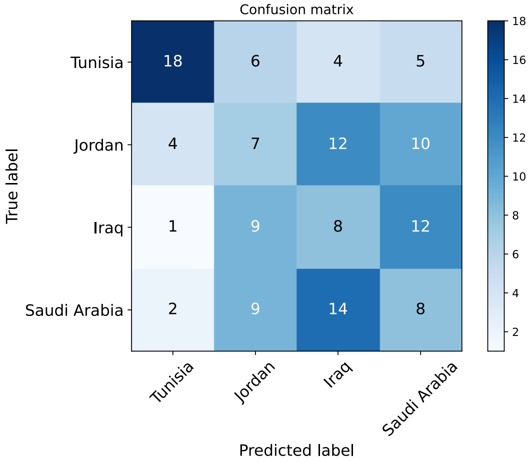


**FIGURE 4.** Confusion matrix for multi-classification using the proposed model on the provided accent dataset.

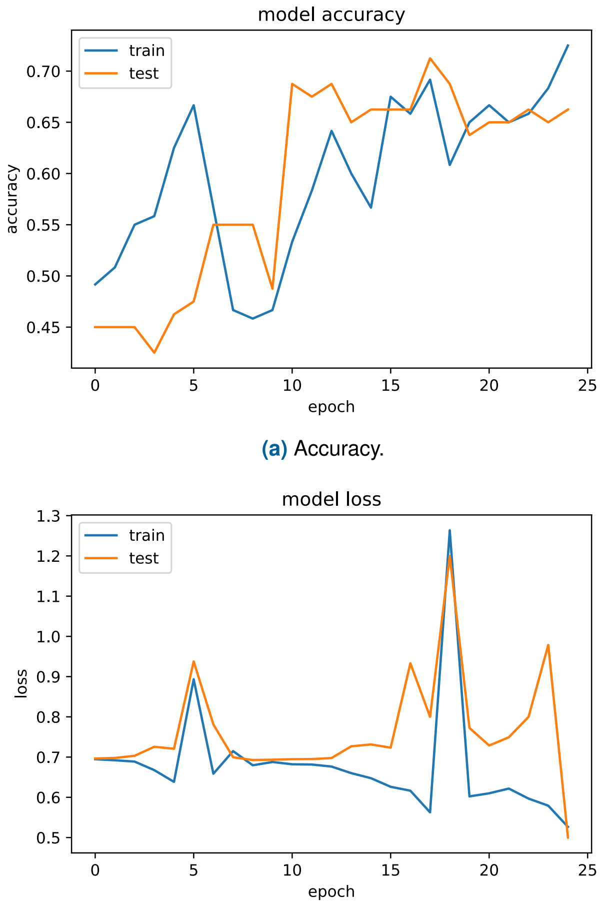
Furthermore, as shown in Table 1, 165 subjects display fluency in reading the stated script, accounting for roughly 42% of the entire number of recordings (391). This poses a challenge for machine learning because fluent speakers of all classes pronounce words and phrases virtually identically. In our investigation, the similarity across classes undermines the distinctiveness essential for meaningful classification. To address this issue, we chose a binary classification strategy, which reduces the similarities between classes. This intentional data split intends to enhance the effectiveness of our study in discerning meaningful differences. The splitting of the data was done as follows:

- Classification of the Tunisian subjects vs the Jordanians.
- Classification of the Tunisian subjects vs the Iraqis.
- Classification of the Tunisian subjects vs the Saudis.
- Classification of the Jordanian subjects vs the Iraqis.
- Classification of the Jordanian subjects vs the Saudis.
- Classification of the Iraqi subjects vs the Saudis.

Since we have some similarities in the accents of the eastern Arab countries (Iraq, Jordan, and Saudi Arabia) [41], particularly when speaking English, we added all of their subjects together forming one class and conducted another experiment to classify Eastern Arab subjects vs Western Arab

**TABLE 4.** Comparison of accuracy (%) between the proposed method and other methods on the provided accent data splits.

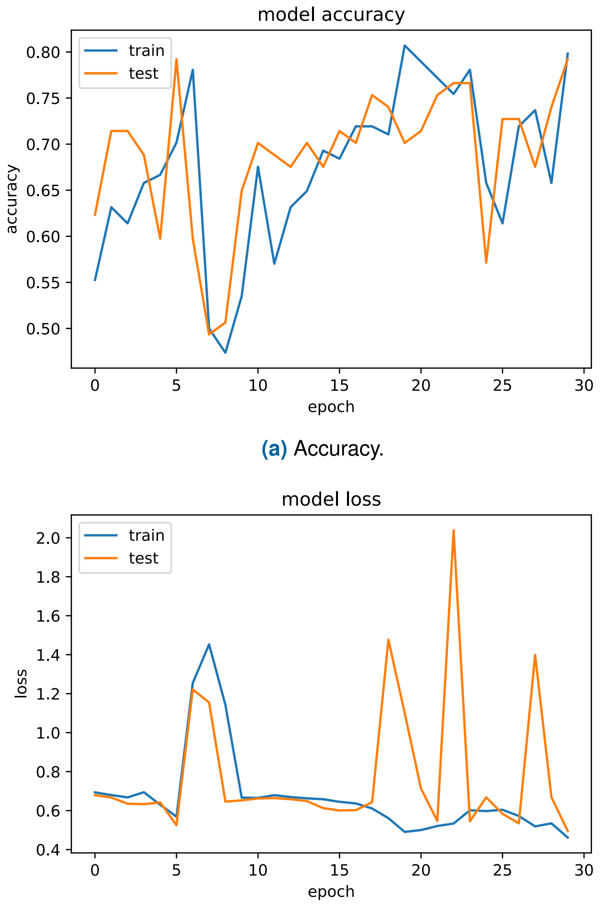
| Experiment | MLP | 1-DCNN | BiLSTM | Proposed LSTM |
|---|---|---|---|---|
| Jordanians vs Iraqis | 45 | 52 | 50 | 54 |
| Jordanians vs Saudis | 52 | 55 | 53 | 61 |
| Iraqis vs Saudis | 47 | 60 | 52 | 57 |
| Tunisians vs Jordanians | 53 | 59 | 62 | 73 |
| Tunisians vs Iraqis | 58 | 64 | 66 | 79 |
| Tunisian vs Saudis | 52 | 55 | 62 | 70 |
| Eastern vs Western Arabs | 58 | 69 | 64 | 72 |



**(a)** Accuracy.

**(b)** Loss function.

**FIGURE 5.** Classification of the Tunisian subjects vs the Jordanians, a) testing and training processes' accuracy outcomes, b) the Loss function of the learning process.



**(a)** Accuracy.

**(b)** Loss function.

**FIGURE 6.** Classification of the Tunisian subjects vs the Iraqis, a) testing and training processes' accuracy outcomes, b) the Loss function of the learning process.

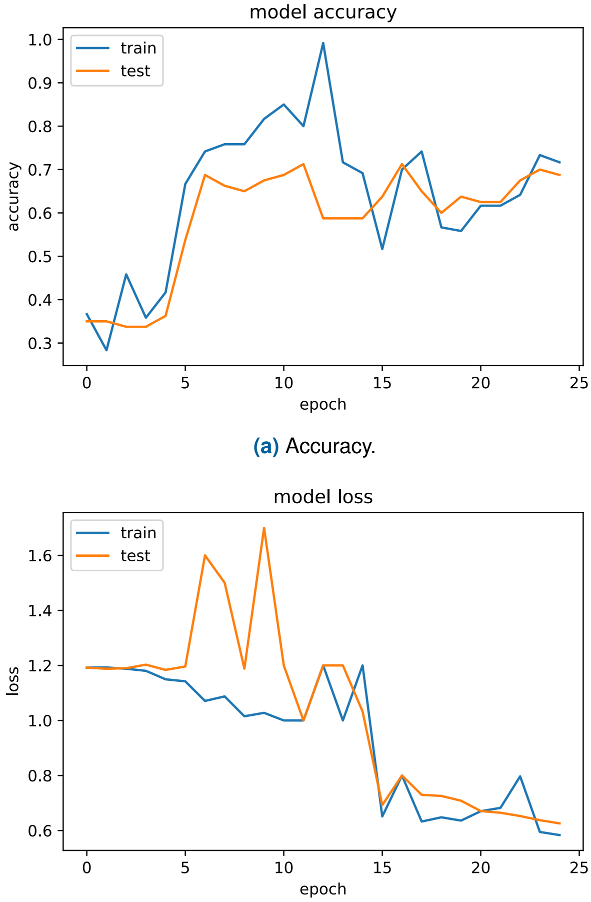
subjects (the Tunisians). The training and testing performance of some of the aforementioned classifications are shown in Figures 5, 6, 7, and 8. The best classification accuracy for each experiment is reported in Table 3.

Another set of experiments is done to compare the proposed method to other methods, including MLP, 1-DCNN, and biLSTM.

These methods require the data in different formats, like MLP it needs the data as one vector. Therefore, we read the data as segments each segment is 5 seconds in length, and the signals are then used to train these classifiers. To avoid having the same speaker in the training and testing sets we split the data in a way to assure the training and testing contains different recordings.

Table 4 shows the results of the comparison between different models to classify the observations in the proposed dataset.

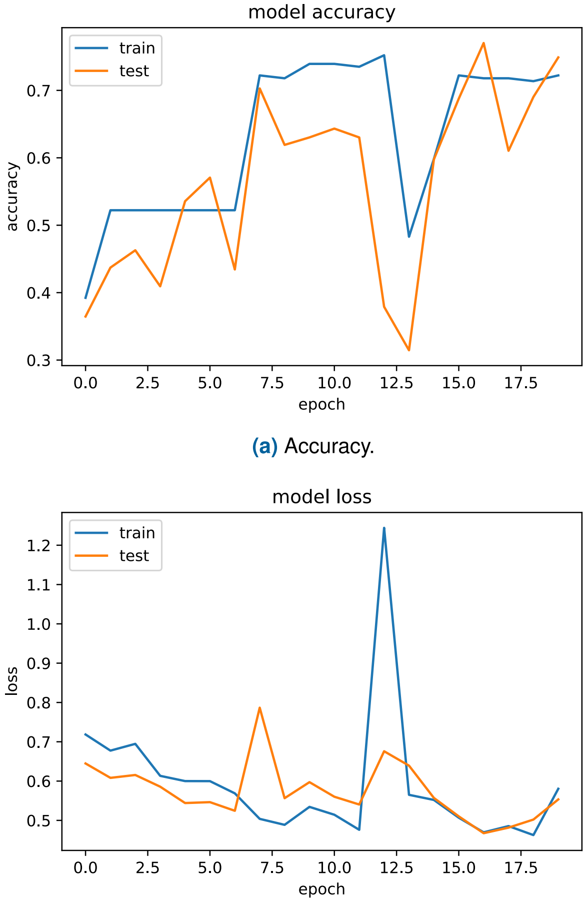
The results presented in Table 4 highlight the superior performance of the proposed method compared to other approaches across various accent data splits. In particular, the proposed LSTM consistently demonstrates a higher accuracy than the alternative methods. It is worth noting



(a) Accuracy.

(b) Loss function.

**FIGURE 7. Classification of the Tunisian subjects vs the Saudis, a) testing and training processes' accuracy outcomes, b) the Loss function of the learning process.**



(a) Accuracy.

(b) Loss function.

**FIGURE 8. Classification of Eastern Arab subjects vs Western Arabs, a) testing and training processes' accuracy outcomes, b) the Loss function of the learning process.**

that the 1-DCNN achieves accuracy results comparable to the proposed method in some instances and on specific data splits. In summary, the results show that the proposed LSTM consistently performs well over multiple experiments, proving its usefulness in accent recognition when compared to the other approaches studied. The findings can help guide method selection decisions in scenarios requiring accent classification tasks.

As can be seen from the results illustrated in Figures 5, 6, 7, and 8, and according to Table 3, the proposed Arabic accent recognition system can recognize speakers of various categories (accents) by up to 79%. This demonstrates that, yes, speakers with an Arabic accent have their way of speaking English, which varies from country to country.

However, the distinctions are not the same. For example, when the system learned to distinguish between Jordanian and Iraqi subjects and then tested their test examples, the accuracy was low (only 54%). This could be attributed to two factors: first, the relatively high number of fluent Jordanian speakers (52 out of 100), and second, the similarity of the Iraqi accent to the Jordanian, given the proximity of both countries.

Similarly, when the system was learned to distinguish between Jordanian and Saudi subjects, as well as Iraqi and Saudi subjects, the accuracy was also low (63% and 57%, respectively). This could also be attributed to the same previous factors, given the three countries' proximity, which fosters some similarity in accent and strategy of learning English, as all three countries have English as a second language, and this could be a third important factor.

When learning the accents of subjects from an eastern Arab country, such as Jordan, Iraq, or Saudi Arabia, and subjects from a western Arab country, such as Tunisia, the proposed system achieves significantly higher accent recognition accuracy. The system also performs better when subjects from Tunisia are used alone in one class and all other eastern subjects are combined in another, as shown in Figure 8.

Obviously, given the large distance between the three eastern countries (Jordan, Iraq, and Saudi Arabia) and Tunisia in the west of the Arab world, which ranges between 3000 and 4000 kilometers. Many studies, such as [42], [43], [44], and [45], show that there is a significant difference in Arabic accents between Eastern and Western Arab countries. According to [46], the Arabic-speaking area can be divided into two major dialects: Eastern (mašriqī) and Western (magribi).

Another interesting factor that might have contributed to the improvement of the system's performance when fed with more distinctive examples is the English language itself, which is used as a second language in countries such as Jordan, Iraq, and Saudi Arabia, and as a third language in Tunisia and other western Arab countries. Perhaps having French as a second language in Tunisia reduces interest in learning English properly, allowing for more distinct ways of pronouncing English words and sentences. Hence, the proposed Arabic accent recognition system performs much better, with accuracy rates of 70%, 73%, 79%, and 72% when learning from (Tunisians and Saudis), (Tunisians and Jordanians), (Tunisians and Iraqis), and (Eastern and Western Arabs) respectively. Having the best performance when learning from subjects residing in one of the most eastern Arab countries (Iraq) and subjects from Tunisia.

In general, the proposed system could not achieve more than 79% accuracy. In addition to the aforementioned reasons, such as the relatively large number of fluent speakers from all countries tested, which hinders the learning process, there is another factor, which is the presence of female speakers in both classes for all experiments (172 females out of 392). It is well known that female voices are more similar to each other than males', providing more similarity between classes, which allows for less accuracy. According to [47], Female participants had higher levels of aspiration noise in the spectral regions corresponding to the third formant, giving female voices a more ''breathy'' quality than male voices.

Although [26] and [27] employed datasets that are different from ours, we compare our results to theirs just to demonstrate that the performance of such systems is not very high and instead to support the idea behind the proposed system, and in this field of research, an accuracy of up to 79% is not considered to be modest.

This study's findings can be summed up as follows:

- A new deep learning method has been presented for analyzing audio speech signals and, in particular, identifying Arabic accents in English speech.
- Due to a lack of accented Arabic speech recognition corpora, a new accented speech dataset was gathered from 391 Arab participants speaking English.
- The proposed system could recognize speakers with various accents by up to 79%. This answers the study's main question, demonstrating that speakers with an Arabic accent have their way of speaking English, which varies by country.

Numerous interesting use cases are possible for our system, especially in any practical context of live speech recognition and translation from voice to text (in English). For example, any intelligent AI-based system involving a real-time automatic interpretation of voice commands is a potential use case. These systems can be useful for executing vocal orders in smart homes, in cars, or on smartphones. It is worth noting that actual live translation systems, even professional commercial ones, failed to recognize non-fluent English voices. Our system will resolve this issue, at least for Arabic speakers of English.

## V. CONCLUSION

This paper investigates how Arab accents affect the English language. To that end, we created a deep machine-learning system for Arabic accent recognition, which was trained on an in-house English speech database containing 391 subjects speaking four Arabic accents from Jordan, Iraq, Saudi Arabia, and Tunisia. All the subjects have spoken one standard English paragraph.

The deep learning model is obtained by training an LSTM neural network to recognize the accent in each sound signal using Mel spectrograms of the English-spoken paragraph. The proposed system could recognize speakers with various accents by up to 79%. This answers the study's main question, demonstrating that speakers with an Arabic accent have their own way of speaking English, which varies by country. We saw such high recognition results when we had two distinctive accents, such as Tunisian vs Iraqi accents.

The limitations of this study include:

- The collected data was extremely difficult to learn due to the significant presence of fluent speakers in all classes.
- The significant presence of female voices in all classes also hinders the learning process.
- Size matters in Deep Learning; despite recording a lengthy paragraph by 391 subjects, our in-house dataset remains small for deep learning.

Our future efforts will be focused on overcoming the limitations listed above: To improve accent learning, fluent speakers must be removed from the speech database, simply because there is no accent information preserved in their recordings. In order to overcome the effect of female voices, our future approach will divide the dataset into two, one for males and one for females, then train the deep learning method on each to produce two models, which are then merged to improve recognition results, as done by [48] and [49]. In addition to collecting more data to increase the size of the dataset for better deep learning. Indeed, we are collecting data from two new countries, Egypt, and Morocco. The actual results will be compared to those of these two countries in future work.

## DATA AVAILABILITY

The data used to support the findings of this study are available from the corresponding author upon request.

## CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

## REFERENCES

[1] S. Montrul and R. Slabakova, "Competence similarities between native and near-native speakers: An investigation of the preterite-imperfect contrast in Spanish," *Stud. 2nd Lang. Acquisition*, vol. 25, no. 3, pp. 351–398, Sep. 2003.

[2] T. M. Derwing and M. J. Rossiter, "ESL learners' perceptions of their pronunciation needs and strategies," *System*, vol. 30, no. 2, pp. 155–166, Jun. 2002.

[3] D. F. Dalton, "Some techniques for teaching pronunciation," *Internet TESL J.*, vol. 3, no. 1, 1997. [Online]. Available: http://iteslj.org/Techniques/Dalton-Pronunciation.html and https://scholar.google.com/scholar?hl=nl&as_sdt=0%2C5&q=Some+Techniques+for+Teaching+Pronunciation+David+F.+Dalton&btnG=

[4] M. Jain, "Error analysis: Source, cause and significance," in *Error Analysis Perspectives on Second Language Acquisition*, 1st ed., J. C. Richards, Ed. London, U.K.: Longman, 1975.

[5] G. F. Simons and C. D. Fennig. (2017). *Ethnologue: Languages of the World, Dallas, Texas: Sil International*. [Online]. Available: http://www.ethnologue.com

[6] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[7] D. H. Y. Tushyeh, "Linguistic problems facing Arab learners of English," *ITL Int. J. Appl. Linguistics*, vols. 111–112, pp. 109–117, Jan. 1996.

[8] S. Winograd, "On computing the discrete Fourier transform," *Math. Comput.*, vol. 32, no. 141, pp. 175–199, 1978.

[9] N. Sturmel and L. Daudet, "Signal reconstruction from STFT magnitude: A state of the art," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, Paris, France, Sep. 2011, pp. 375–386.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[11] A. Ahmed and I. Lenchuk, "Testing the interface hypothesis: The acquisition of English indirect questions by L1 speakers of omani Arabic," *Heliyon*, vol. 8, no. 1, Jan. 2022, Art. no. e08752.

[12] M. O. Elfahal, M. Mustafa, M. E. Mustafa, and R. A. Saeed, "A framework for Sudanese Arabic—English mixed speech processing," in *Proc. Int. Conf. Comput. Inf. Technol. (ICCIT)*, Sep. 2020, pp. 1–6.

[13] I. Ababneh, "English pronunciation errors made by Saudi students," *Eur. Sci. J., ESJ*, vol. 14, no. 2, pp. 244–261, Jan. 2018.

[14] D. Jiao, V. Watson, S. G.-J. Wong, K. Gnevsheva, and J. S. Nixon, "Age estimation in foreign-accented speech by non-native speakers of English," *Speech Commun.*, vol. 106, pp. 118–126, Jan. 2019.

[15] M. Al-Shumari and G. Bella, "Online English vocabulary learning on different systems for non-english speakers," in *Proc. ELMAR*, Sep. 2014, pp. 1–4.

[16] S. F. Jiang, B.-C. Yan, T.-H. Lo, F.-A. Chao, and B. Chen, "Towards robust mispronunciation detection and diagnosis for L2 English learners with accent-modulating methods," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 1065–1070.

[17] R. Kethireddy, S. R. Kadiri, and S. V. Gangashetty, "Deep neural architectures for dialect classification with single frequency filtering and zero-time windowing feature representations," *J. Acoust. Soc. Amer.*, vol. 151, no. 2, pp. 1077–1092, Feb. 2022.

[18] M. Ibrahim, M. Nayeem, and S. Al Arabi, "Predicting regional accents of Bengali language using deep learning," Ph.D. dissertation, Dept. Comput. Sci. Eng., Brac Univ., Dhaka, Bangladesh, 2021. [Online]. Available: http://hdl.handle.net/10361/15872

[19] Y. Lan and T. Xie, "L2 accent and intelligibility by Chinese L2 speakers of English," in *Proc. 24th Conf. Oriental COCOSDA Int. Committee Co-Ordination Standardisation Speech Databases Assessment Techn.*, Nov. 2021, pp. 82–87.

[20] T. Shevchenko, "Kuanyshbay Darkhan Nurgazyuly development of methods, algorithms of machine learning and mobile applications for Kazakh speech recognition," Ph.D. dissertation, Dept. Comput. Sci., Faculty Eng. Natural Sci., Nat. Univ., San Diego, CA, USA, 2021.

[21] S. Khanal, M. T. Johnson, M. Soleymanpour, and N. Bozorg, "Mispronunciation detection and diagnosis for Mandarin accented English speech," in *Proc. Int. Conf. Speech Technol. Human Computer Dialogue (SpeD)*, Oct. 2021, pp. 62–67.

[22] R. Kethireddy, S. R. Kadiri, and S. V. Gangashetty, "Learning filterbanks from raw waveform for accent classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–6.

[23] R. Kethireddy, S. R. Kadiri, and S. V. Gangashetty, "Exploration of temporal dynamics of frequency domain linear prediction cepstral coefficients for dialect classification," *Appl. Acoust.*, vol. 188, Jan. 2022, Art. no. 108553.

[24] R. Kethireddy, S. R. Kadiri, P. Alku, and S. V. Gangashetty, "Mel-weighted single frequency filtering spectrogram for dialect identification," *IEEE Access*, vol. 8, pp. 174871–174879, 2020.

[25] R. Kethireddy, S. R. Kadiri, S. Kesiraju, and S. V. Gangashetty, "Zero-time windowing cepstral coefficients for dialect classification," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey), Int. Speech Commun. Assoc. (ISCA)*, 2020, pp. 32–38, doi: 10.21437/Odyssey.2020-5.

[26] S. Mnasri and M. Habbash, "Study of the influence of Arabic mother tongue on the English language using a hybrid artificial intelligence method," *Interact. Learn. Environments*, vol. 31, no. 9, pp. 5568–5581, Dec. 2023.

[27] A. Ali, N. Dehak, P. Cardinal, S. Khurana, S. H. Yella, J. Glass, P. Bell, and S. Renals, "Automatic dialect detection in Arabic broadcast speech," in *Proc. Interspeech*, Sep. 2016, pp. 2934–2938.

[28] J. Opadere, "Improvement of customer baseline calculation methodologies of demand response using maximal overlap discrete wavelet packet transform," Ph.D. dissertation, Dept. Elect. Comput. Eng., Univ. North Carolina Charlotte, Charlotte, NC, USA, 2020.

[29] S. Gupta, A. T. Patil, M. Purohit, M. Parmar, M. Patel, H. A. Patil, and R. C. Guido, "Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments," *Neural Netw.*, vol. 139, pp. 105–117, Jul. 2021.

[30] U. S. Shanthamallu, S. Rao, A. Dixit, V. S. Narayanaswamy, J. Fan, and A. Spanias, "Introducing machine learning in undergraduate DSP classes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7655–7659.

[31] M. H. Tanveer, H. Zhu, W. Ahmed, A. Thomas, B. M. Imran, and M. Salman, "Mel-spectrogram and deep CNN based representation learning from bio-sonar implementation on UAVs," in *Proc. Int. Conf. Comput., Control Robot. (ICCCR)*, Jan. 2021, pp. 220–224.

[32] S. Weinberger. (2015). *Speech Accent Archive. George Mason University*. Accessed: Mar. 2, 2024. [Online]. Available: http://accent.gmu.edu

[33] S. H. Weinberger and S. Kunath, "Towards a typology of English accents," in *Proc. AACL*, vol. 104, 2009.

[34] M. Jekiel, "L2 rhythm production and musical rhythm perception in advanced learners of English," *Poznan Stud. Contemp. Linguistics*, vol. 58, no. 2, pp. 315–340, Jun. 2022.

[35] J. Lee, S. Dhar, R. Abel, R. Banakis, E. Grolley, J. Lee, S. Zecker, and J. Siegel, "Behavioral hearing thresholds between 0.125 and 20 kHz using depth-compensated ear simulator calibration," *Ear Hearing*, vol. 33, no. 3, pp. 315–329, 2012.

[36] S. Caro-Alvaro, A. A. Alkasasbeh, E. Garcia-Lopez, A. Garcia-Cabot, G. Rozinaj, and G. Ghinea, "Examining potential of scents for enhancement of user performance with mobile apps," *Mobile Inf. Syst.*, vol. 2022, pp. 1–11, Mar. 2022.

[37] A. B. A. Hassanat, V. B. S. Prasath, M. Al-kasassbeh, A. S. Tarawneh, and A. J. Al-shamailh, "Magnetic energy-based feature extraction for low-quality fingerprint images," *Signal, Image Video Process.*, vol. 12, no. 8, pp. 1471–1478, Nov. 2018.

[38] A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh, and A. Almuhaimeed, "Stop oversampling for class imbalance learning: A review," *IEEE Access*, vol. 10, pp. 47643–47660, 2022.

[39] E. Al-Mahadeen, M. Alghamdi, A. S. Tarawneh, M. A. Alrowaily, M. Alrashidi, I. S. Alkhazi, A. Mbaidin, A. A. Alkasasbeh, M. A. Abbadi, and A. B. Hassanat, "Smartphone user identification/authentication using accelerometer and gyroscope data," *Sustainability*, vol. 15, no. 13, p. 10456, Jul. 2023.

[40] A. B. Hassanat, H. N. Ali, A. S. Tarawneh, M. Alrashidi, M. Alghamdi, G. A. Altarawneh, and M. A. Abbadi, "Magnetic force classifier: A novel method for big data classification," *IEEE Access*, vol. 10, pp. 12592–12606, 2022.

[41] W. F. Alshammari, "Numeral form selection and accommodation in Gulf Pidgin Arabic," *Lang., Interact. Acquisition*, vol. 13, no. 1, pp. 29–62, Aug. 2022.

[42] S. Ghazali, R. Hamdi, and M. Barkat, "Speech rhythm variation in Arabic dialects," in *Proc. Int. Conf. Speech Prosody*, Apr. 2002.

[43] M. Barkat, J. Ohala, and F. Pellegrino, "Prosody as a distinctive feature for the discrimination of Arabic dialects," in *Proc. 6th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Budapest, Hungary, Sep. 1999, pp. 395–398.

[44] S. Al Yaari, M. Alkhunayn, M. Al Yaari, A. Al Yaari, A. Al Yaari, A. Al Yaari, S. Al Yaari, and F. Eissa, "On the weightlessness of vowel lengthening: Insights from Arabic dialect of Yemen and contribution to psychoneurolinguistics," *Int. Rev. Social Sci.*, vol. 10, no. 6, pp. 1–116, 2022.

[45] C. Solimando, "Language tests in the arabic-speaking world: Between ideology and language policy," *Lingue Culture Mediazioni Lang. Cultures Mediation*, vol. 8, no. 2, pp. 13–31, Feb. 2022.

[46] K. Versteegh, A. Elgibali, and A. Zaborski, "Encyclopedia of Arabic language and linguistics," *Afr. Stud.*, vol. 8, pp. 57–263, Jul. 2009.

[47] J. Van Borsel, J. Janssens, and M. De Bodt, "Breathiness as a feminine voice characteristic: A perceptual approach," *J. Voice*, vol. 23, no. 3, pp. 291–294, May 2009.

[48] V. B. S. Prasath, I. S. Alkhazi, M. Alghamdi, M. Alanazi, H. Alharbi, M. Alrashidi, A. S. Tarawneh, A. A. Albustanji, and A. B. A. Hassanat, "DeepVeil: Deep learning for identification of face, gender, expression recognition under veiled conditions," *Int. J. Biometrics*, vol. 14, nos. 3–4, p. 453, 2022.

[49] A. S. Tarawneh, A. B. Hassanat, E. Alkafaween, B. Sarayrah, S. Mnasri, G. A. Altarawneh, M. Alrashidi, M. Alghamdi, and A. Almuhaimeed, "DeepKnuckle: Deep learning for finger knuckle print recognition," *Electronics*, vol. 11, no. 4, p. 513, Feb. 2022.

**MANSOOR HABBASH** occupied the position of the Dean of the Community College, University of Tabuk, from 2018 to 2021, where he is currently a Professor in English linguistics. He is also have different responsibilities with the University of Tabuk. He specializes in language studies. He conducted different projects and research works regarding applied linguistics, especially in: undertaking a needs analysis with computer students, critical analysis of the status of the English language in Saudi Arabia, discourse of global English and its representation in the Saudi context, postmodernism and critical theory, critical pedagogy and critical applied linguistics, and curriculum and syllabus design and language program evaluation.

**SAMI MNASRI** received the Ph.D. degree in computer science from the University of Toulouse. He is currently a Computer Science Assistant Professor with the Applied College, University of Tabuk. He has published several high-impacted research studies regarding the utility of using evolutionary optimization strategies, such as genetic algorithms, particle swarms, and ant colony optimization on the resolution of real-world complex problems in the context of IoT networks. He also did several studies on multi-agent systems and artificial intelligence. His current research interests include machine learning, deep learning, and blockchain. He is the Organizing Co-Chair of the IEEE International IINTEC Conference and the WSDWSN International Workshop. He was a Speaker of Saudiiot, IINTEC, and HIS Conferences.

**MANSOOR ALGHAMDI** was a Supervisor of the Quality Assurance Unit, Deanship of Quality and Development. Since 2019, he has been the Vice Dean of Information Technology for Quality and Development. He has been an Assistant Professor and the Dean of the Applied College, University of Tabuk, since 2021. His work in his Ph.D. focused on simulating and modeling the computer worms in peer-to-peer networks. His current research interests include investigating the utilization of machine learning, deep learning, and blockchain techniques in resolving networking and smart cities issues. His research interests include vision and network science, virtual reality, augmented reality, the Internet of Things, robotics, big data, artificial intelligence, and wireless sensor networks.

**MALEK ALRASHIDI** received the B.Sc. degree in computer science from Taibah University, Madinah, Saudi Arabia, in 2008, the M.Sc. degree in computer science from Newcastle University, Newcastle, U.K., in 2011, and the Ph.D. degree in computer science from the University of Essex, Colchester, U.K., in 2017. He is currently the Head of the Computer Department and the Vice Dean of the Community College for Academic Affairs, Applied College, University of Tabuk, Saudi Arabia. His research interests include artificial intelligence, the IoT, 3D user interface, augmented reality, and human–computer interaction. His current research interests include virtual reality, augmented reality, the Internet of Things, robotics, big data, AI, and WSN. He is a member of the Technical Program Committee at Computer and Electronic Engineering Conference, University of Essex. He was a member of the Organizing Committee of Realistic Workshop Smart Environments, Shanghai, China, in June 2014.

**AHMAD S. TARAWNEH** was born in Kerak, Jordan. He received the B.S. and M.S. degrees in computer science from Mutah University, in 2013 and 2015, respectively, and the Ph.D. degree in computer science from Eötvös Loránd University (ELTE), Hungary, Budapest. He worked on the project EFOP (image and video processing), which is sponsored by the Hungarian government and co-financed by the European Social Fund. Currently, he is an Assistant Professor with the Department of Data Science and AI, Mutah University. His main research interests include computer vision, deep learning, machine learning, and data mining.

**ABDULLAH GUMAIR** is currently pursuing the master's degree with the College of Computers & Information Technology, University of Tabuk, Saudi Arabia. His main research interests include the Internet of Things and pattern recognition.

**AHMAD B. HASSANAT** (Senior Member, IEEE) was born in Jordan. He received the B.S. degree in computer science from Mutah University, Jordan, in 1995, the M.S. degree in computer science from Al-Albayt University, Jordan, in 2004, and the Ph.D. degree in computer science from the University of Buckingham, U.K., in 2010. Since 2010, he has been a Faculty Member with the Faculty of Information Technology, Mutah University. His primary research interests include computer vision, machine learning, big data, and pattern recognition.