IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 12, 2022, accepted May 17, 2022, date of publication May 23, 2022, date of current version June 3, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3177191

Deep Investigation of the Recent Advances in Dialectal Arabic Speech Recognition

HAMZAH A. ALSAYADI[1,2], **ABDELAZIZ A. ABDELHAMID**[1,5] **ISLAM HEGAZY**[1], **BANDAR ALOTAIBI**[3,4], **(Member, IEEE), AND ZAKI T. FAYED**[1]

[1]Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt
[2]Computer Science Department, Faculty of Sciences, Ibb University, Ibb, Yemen
[3]Information Technology Department, University of Tabuk, Tabuk 71491, Saudi Arabia
[4]Sensor Networks and Cellular Systems Research Center, University of Tabuk, Tabuk 71491, Saudi Arabia
[5]College of Computing and Information Technology, Shaqra University, Shaqra 11961, Saudi Arabia

Corresponding authors: Hamzah A. Alsayadi (hamzah.sayadi@cis.asu.edu.eg) and Abdelaziz A. Abdelhamid (abdelaziz@cis.asu.edu.eg)

This work was supported in part by the Sensor Networks and Cellular Systems (SNCS) Research Center under Grant 1443-001.

ABSTRACT Speech recognition systems play an important role in human–machine interactions. Many systems exist for Arabic speech, however, there are limited systems for dialectal Arabic speech. The Arabic language comprises many properties, some of which are ideal for building automatic speech recognition systems such as syntax and phonology, while other properties are unsuitable for developing speech systems. Importantly, most data are in non-diacritized form, vary in dialect, and contain morphological complexity. Moreover, the Arabic dialects lack a standard structure. In this paper, we highlighted the works and frameworks that have been developed in the last fourteen years of dialectal Arabic speech recognition systems. The paper also presented an analysis and evaluation for several studies using different approaches and techniques. The main goal of this paper is to compare and discuss different studies in dialectal Arabic speech systems including several criteria such as techniques, datasets, evaluation metrics, and dialect types. The study also includes a description of some techniques used in all steps of the dialectal Arabic speech system such as hidden Markov models (HMM), convolutional neural network (CNN), and deep neural network (DNN). In addition, we introduced the challenges and problems of dialectal Arabic speech recognition systems. Overall, more studies are required to obtain a more accurate speech system for dialectal Arabic.

INDEX TERMS Speech recognition, Arabic dialect, dialectal Arabic ASR, acoustic modeling.

I. INTRODUCTION

Automatic speech recognition (ASR) is one of the earliest tasks in artificial intelligence (AI) research, which is used to convert speech waves or signals to a mapping words (units) sequence using an appropriate algorithm [1]. ASR has a wide area of IT applications: employing a range of IT solutions and applications for civil areas and industry, human–computer interactions (HCI), voice applications, automatic language translation, and many via-voice systems [2]. Unlike other languages, there is limited research on Arabic speech recognition systems. Arabic speech recognition systems are a difficult task, that is due to many reasons: the data sparseness of the language, lexical variety, number of several dialects spoken in the world, and the predominance of non-diacritized text material. Moreover, the Arabic language is a morphologically complex language. However, the Arabic language is very rich in vocabulary [3]. Thus, large-vocabulary ASR for Arabic also represents several challenges for speech research. Over the past decade, researchers have been greatly interested in building robust Arabic automatic speech recognition (AASR) systems [4], [5]. The Arabic language has a set of special symbols (marks)-called diacritics (Arabic harakat)-that are placed above the main symbols (letters) [6]. These diacritics represent sounds similar to vowels in English and tones in Chinese. These letter sounds are important for understanding the meaning of words and sentences and represent another challenge for Arabic ASR.

In addition, building acoustic models for dialectal Arabic ASR is challenging. In these systems, training the model requires the appropriate Arabic dialect. Therefore, developing Arabic dialectal ASR has several challenges. Lacking

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

enough training data: A large data set should be collected to obtain a good and accurate model. Unfortunately, collecting dialect training data is a difficult task compared to other modern standard Arabic (MSA) versions and languages. The problems mainly correspond to building an accurate transcription for dialect versions. The variety of dialects is a challenge since the Arabic dialect has many different forms (Egyptian, Levantine, Iraqi, Gulf, etc.). Furthermore, each village sometimes has a different dialect form. In addition, collecting a pronunciation dictionary that includes whole dialectal words is also immensely challenging. The Arabic dialect lacks standard orthography for writing and is mostly spoken language. The diacritization for dialectal Arabic is far more challenging than MSA since it requires a dialectal Arabic morphological analyzer to generate various diacritization forms. Using context-based forms, the diacritization also requires a robust language model for dialectal Arabic which is currently unavailable. Moreover, the dialectal Arabic diacritization-using automatic alignment against the audio signal-is also difficult due to the larger set of vowels [7]. Therefore, the Arabic dialect does not include diacritics. Thus, this leads us to build an inaccurate and less predictive language model. Furthermore, the high morphological complexity and high degree of morphological complexity, during decoding, lead to high out-of-vocabulary rates and larger search spaces [7], [8].

This paper aims to highlight the last achievements in dialectal Arabic ASR. We a comprehensive overview for dialectal Arabic ASR systems including different criteria such as techniques, datasets, evaluation metrics, and dialect types. The paper includes an analysis and discussion several studies in this domain. The description of components and methods of dialectal Arabic ASR are presented. In addition, the study also presents some current challenges and difficulties facing the system developers. We also introduce the knowledge of techniques that are used in these systems and investigate the used approaches and open-source data sets of several Arabic dialects.

The rest of the paper is organized as follows. Section 2 introduces the research methodology. In Section 3, we present a literature review for dialectal Arabic ASR. In Section 4, the main steps for ASR are described. Section 5 presents the discussion and challenges. Finally, Section 6 includes the conclusion and suggestions.

II. RESEARCH METHODOLOGY

The research methodology includes a number of steps as shown in Figure 1. We used Google and Google Scholar to search for studies in the dialectal Arabic ASR field. Some of the keywords and strings are used to find studies and manuscripts related to our research topic. These keywords and strings include: ‘‘dialectal Arabic automatic speech system’’, ‘‘dialect Arabic automatic speech system’’, ‘‘dialectal Arabic ASR’’, ‘‘automatic speech system for dialect Arabic’’, ‘‘dialect Arabic’’, ‘‘Arabic automatic speech system’’, ‘‘Arabic ASR’’, and ‘‘English language’’. If studies include any keywords or strings in their content, we simply filter these studies to select the suitable manuscript for dialectal Arabic ASR. We also use the citations of some manuscripts to obtain other manuscripts regarding our topic. The date of the initial collection is between 2005 and 2022. The initial number of collected manuscripts is 130 papers. Then, the range of date is reduced to 2009–2022 and the article type is selected resulting in 76 papers. Finally, we manually selected the studies that are related to dialectal Arabic ASR based on some criteria such as: (1) Studies that presented the speech recognition (speech-to-text) systems; (2) studies that include systems for pure dialects Arabic, i.e., the developed systems only used pure Arabic dialects for training and evaluation; (3) studies that used the Arabic dialects as part of the training and evaluation data; (4) studies that utilized the Arabic dialects for evaluation; (5) studies that utilized the Arabic for dialects adaption and testing the acoustic model; (6) studies that used the Arabic dialects for speech code-switching. Thus, the studies or papers containing information, descriptions, or related works for dialectal Arabic ASR were excluded—i.e., studies that do not include investigation and results of dialectal Arabic ASR were not taken into consideration. After the manual selection process, 35 studies were reviewed and analyzed in this review according to criteria such as techniques, approaches, datasets, evaluation metrics, dialect types, and well-known publishers.

III. LITERATURE REVIEW

In this paper, 35 studies for dialectal Arabic ASR published in different journals and conferences over the last 13 years were introduced. Most studies were developed using machine learning methods. We present several studies for various Arabic dialect types.

Soltau *et al.* [9] presented a description for Arabic broadcast transcription evaluation using several techniques. They used a large vocabulary and cross-adaptation for two acoustic models (unvowelized and vowelized) to enhance the performance. Hidden Markov models (HMMs) were used for building acoustic models with a mixture of diagonal-covariance Gaussian densities. Feature space maximum likelihood linear regression (FMLLR), maximum likelihood linear regression (MLLR), and feature minimum phone error (fMPE) were utilized as discriminative training techniques. The global autonomous language exploitation (GALE) Phase 2 and Arabic Gigaword corpus were used for training and evaluating the acoustic and language models. For dialect-specific acoustic modeling, experiments are reported as a decision tree depending on the dialect-dependent questions. The obtained results included 25.9% for the regular tree and 24.7% for the dialect tree.

Elmahdy *et al.* [10] introduced a new multilingual system for dialectal ASR. The HMM-based technique was used for training with MLLR, maximum a posteriori (MAP), and vocal tract length normalization (VTLN) as adaptation techniques. The acoustic models used the news broadcast corpus of MSA for decoding Egyptian colloquial Arabic (ECA).

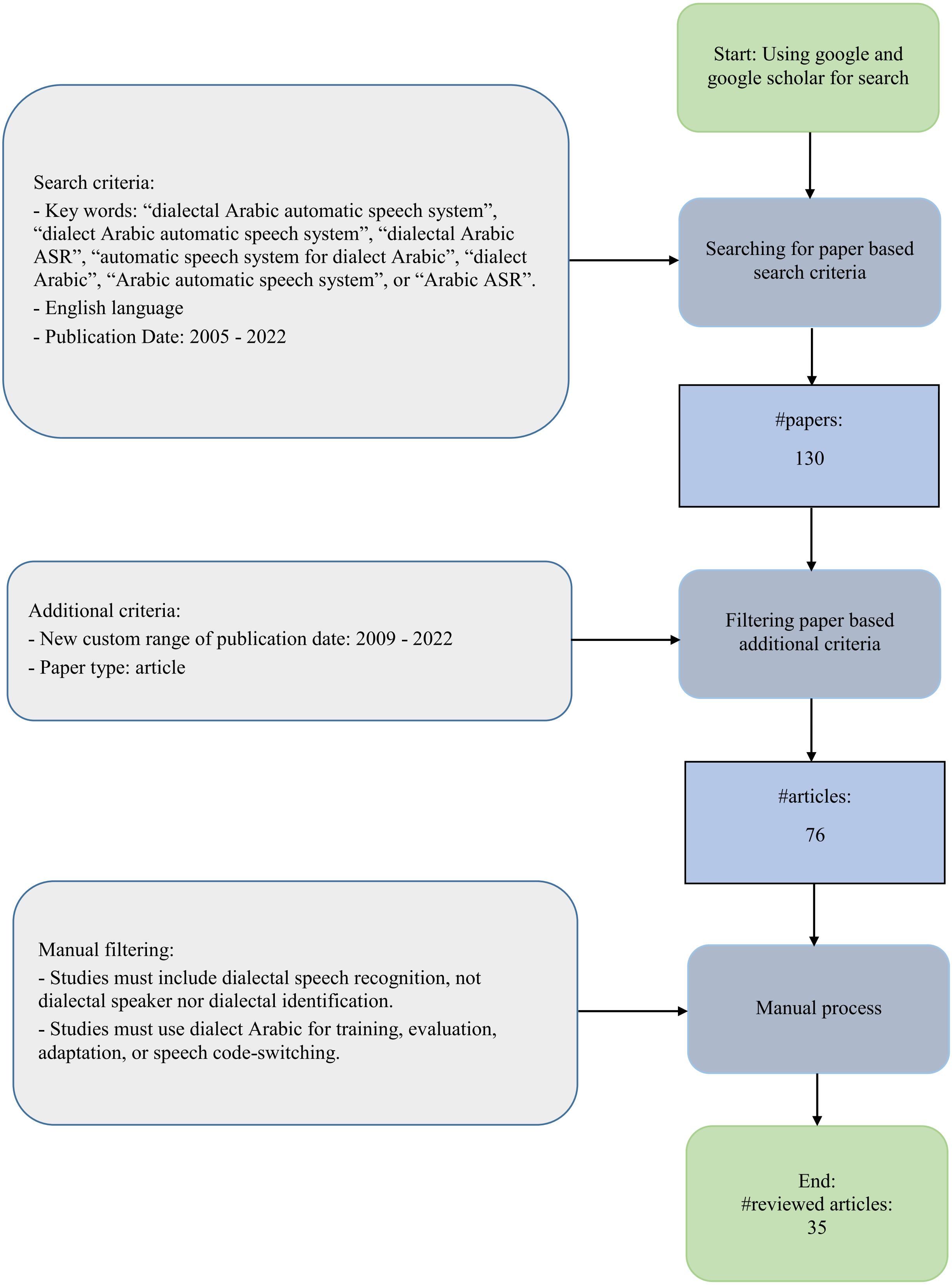

Start: Using google and google scholar for search
Search criteria:
- Key words: “dialectal Arabic automatic speech system”, “dialect Arabic automatic speech system”, “dialectal Arabic ASR”, “automatic speech system for dialect Arabic”, “dialect Arabic”, “Arabic automatic speech system”, or “Arabic ASR”.
- English language
- Publication Date: 2005 - 2022
Searching for paper based search criteria
#papers:
130
Additional criteria:
- New custom range of publication date: 2009 - 2022
- Paper type: article
Filtering paper based additional criteria
#articles:
76
Manual filtering:
- Studies must include dialectal speech recognition, not dialectal speaker nor dialectal identification.
- Studies must use dialect Arabic for training, evaluation, adaptation, or speech code-switching.
Manual process
End:
#reviewed articles:
35

FIGURE 1. Overview of filtering process of the reviewed studies.

The authors collected the ECA connected digits data for evaluating their model. An accuracy rate of 99.34% was reached.

Al-Haj *et al.* [11] proposed a model to recognize the dialectal Iraqi-Arabic. Pronunciation modeling was used for investigating the dialectal Iraqi-Arabic. This acoustic model is a combined model using HMM-based, sub-phonetically tied, and semi-continuous. The Mel frequency cepstral coefficient (MFCC) and approximations of the first and second derivatives were used for features extraction with 42 dimensional coefficients. The models were trained for 450 hours of the Iraqi-Arabic dataset. The results were evaluated by two versions of evaluation data; the best results were 35.84% and 33.30% using multi-pronunciation with estimated weights.

Selouani and Boudraa [12] created the dialectal Algerian database (known as the Algerian Arabic speech database (ALGASD)). This database includes 300 Algerian native speakers. For training and evaluating this data, the authors built ASR using the hidden Markov model toolkit (HTK) and MFCCs for features extraction. In experiments, the test data consisted of 157 sentences for evaluating the system. The results achieved an accuracy rate of 91.65%.

Elmahdy *et al.* [13] proposed an ASR system for ECA depending on the benefit of MSA resources. Cross-lingual acoustic modeling was suggested using the Gaussian mixture model (GMM) and HMM. MLLR and MAP were used for adapting the acoustic model. The authors investigated phoneme-based and graphene-based acoustic modeling to adapt the MSA model using spelling variants. This adaptation was used to select the correct ECA spelling. The results showed a word error rate (WER) of 35.00% with MLLR, MAP, and spelling variants.

Saon *et al.* [14] introduced a description for the Arabic broadcast transcription using a mixture of GALE, FBIS, and topic detection and tracking (TDT-4) audio. Subspace Gaussian mixture models (SGMM) were utilized to train the acoustic model and neural network acoustics were utilized to train the language model (LM). They used modified Kneser–Ney smoothing for enhancing LM. MLLR and fMLLR are used to estimate the acoustic models using speaker-independent (SI). The best WER result was 9.10% with the language model. Elmahdy *et al.* [15] suggested the dialectal Arabic speech transcription system using the Arabic chat alphabet (ACA). GMM-HMM was used to train acoustic models depending on phoneme-based and grapheme-based. Kneser–Ney smoothing was used to train a bi-gram LM. The ECA corpus with collected ACA data was used for training. The best WER of this work was found to be 13.40%.

Huang and Hasegawa-Johnson [16] presented an Arabic ASR system to classify phones based on West point MSA with Babylon Levantine Arabic corpus. They proposed cross-dialectal GMM as a training method to train the acoustic model and used transfer learning to transfer MSA data into dialectal Levantine Arabic.

Biadsy *et al.* [17] built Google's Arabic voice search system for multiple Arabic dialects and made a compression between each. These dialectal languages were Egypt (EG), Jordan (JO), Lebanon (LB), Saudi Arabia (SA), and the United Arab Emirates (AE). They used the standard 3-HMM state for training the acoustic model and boosted maximum mutual information (MMI) as discriminative training techniques. In features extraction, linear discriminant analysis (LDA) was used as an adaptation method. The language model was trained as 5-gram backoff LMs using entropy pruning and Katz smoothing. The results for all used dialectal language were 27.7, 28.7, 24.6, 18.5, and 24.2 for AE, SA, EG, JO, and LB, respectively.

Almeman and Lee [18] proposed the Arabic ASR system for recognizing the MSA, Egyptian, Gulf, and Levantine dialects. This work presented compression between different Arabic dialect languages. The CMU Sphinx framework was used for training the acoustic model. The best WERs achieved were 13.7%, 10.00%, 17.00%, 15.10%, and 16.30% for multi-dialect, MSA, Gulf, Levantine, and Egyptian, respectively.

Masmoudi *et al.* [19] presented a novel Tunisian Arabic corpus and dictionary for ASR which was coined Tunisian Arabic railway interaction corpus (TARIC). The Tunisian graphemes were converted into the corresponding phonemes using rule-based tools. Moreover, the authors built a tool for this rule-based depending on a set of graphemes, phonemes, the lexicon of exceptions, and phonetic rules. Two types of corpora were used for evaluating the performance of ruled-based tools and pronunciation dictionaries. The results showed a WER of around 9%.

Ali *et al.* [20] developed an under-resourced Egyptian ASR system and presented the results for this system. GMM, SGMM, and deep neural network (DNN) models were employed for training the acoustic model using the KALDI toolkit. Minimum phone error (MPE) and bMMI were used as discriminative training to adapt the acoustic model. The SRILM toolkit was utilized to build LM with Kneser–Ney smoothing. The standard 13-dimensional cepstral mean-variance normalized (CMVN) MFCC was used to extract features. The training and evaluation were conducted using 10 hours dataset. An optimal WER of 44.71% was obtained for EG grapheme.

Elmahdy *et al.* [21] proposed the dialectal Arabic ASR system for the Qatari Arabic (QA) dialect based on an under-resourced Arabic dialect. The GMM-HMM architecture was used to train the model using the KALDI toolkit. The transfer learning technique was proposed to transfer MSA-based into the Qatari dialect. In the transfer learning method, they used different processes to increase the accuracy such as orthographic normalization, phone mapping, data pooling, acoustic model adaptation, and combined model. The backoff tri-gram model is built with Kneser–Ney smoothing. An optimal WER of 64.4%—for the evaluation—was obtained with a combination of data pooling methods, adaptation methods, and Lattice minimum Bayes risk (MBR) decoding.

Wray *et al.* [22] presented a study to assess the quality control in crowdsourcing transcriptions. The Arabic ASR system was built on MSA and dialect Arabic data. Moreover, the dialect data was used to evaluate the quality of transcription with an Edit distance algorithm. In Egyptian and North African dialects, the transcription error was reduced by 1.0% for Egyptian data and 4.0% for North African data.

Ali *et al.* [23] proposed a method for measuring accuracy of the ASR system. They presented a new approach to report the accuracy of the ASR system in a non-standard orthographic language known as the multi-reference word error rate (MRWER). The grapheme-based approach was used for building an acoustic model using sequential DNN. In the experiment, an MRWER of 53% was obtained, and WERs of 76.4% and 80.9% were reported.

Khurana and Ali [24] presented a description for the dialectal Arabic multi-genre broadcast (MGB-2) challenge, evaluated based on 1,200 hours of speech audio. They proposed an LF-MMI modeling framework for building the system. The system was trained separately using long short-term memory (LSTM), BLSTM, and TDNN techniques. A recurrent neural network (RNN) was used for building the 4-gram language model with MaxEnt connections (RNNME) by the RNNLM toolkit. The features were transformed using adaptation techniques such as LDA, MLLT, and fMLLR. The KALDI toolkit was used for building all trained models. Finally, the three models were combined into one model that achieved a WER of 14.2%.

Amazouz *et al.* [25] introduced a study for the effectiveness of code-switching (CS) in an ASR system. CS in French/Algerian Arabic was proposed for comparing the quantity of CS that occurred in dialectal Arabic and switching to French. They built the acoustic-phonetic model based on collected Maghrebian broadcasts news data including Algerian, Moroccan, and Tunisian dialects. The results showed that the Algerian dialect had a better CS rate than Tunisian and Moroccan.

Masmoudi *et al.* [26] proposed a framework for developing an ASR system based on the Tunisian dialect. They sought to summarize the linguistic characteristics, such as phonological, morphological, and syntactic, of the Tunisian dialect. This work introduced grapheme-to-phoneme (G2P) conversion using the ruled-based technique. An accuracy of 22.60% was obtained as WER.

Menacer *et al.* [27] developed an ASR system for MSA and Algerian dialect known as Arabic Loria ASR (ALASR). The DNN-HMM technique was utilized to build the acoustic model, and the LM was built by a classical n-gram. The sMBR (state-level minimum Bayes risk) criterion was applied for adapting the training. The Kaldi toolkit was utilized for building the acoustic model. The Nemlar, Gigaword Arabic, and NetDC corpora were used for training and testing the model. WERs of 14.02%, 89%, and 65.45% were obtained for MSA, Algerian dialect, and the combined data (MSA and Algerian), respectively.

Ali *et al.* [28] presented a detailed description of the Arabic MGB-3 challenge. The MGB-3 was used for evaluating the Arabic ASR system. MGB-3 consists of 16 hours of Egyptian dialect that are collected from talk show programs on YouTube. The system was trained using LSTM, BLSTM, and TDNN techniques. The lexical and i-vector bottleneck features were extracted for use in this system. The system was evaluated using MGB-3 testing with an average WER of 37.5%.

Ali *et al.* [29] introduced a study assessing the effectiveness of the ASR on dialectal Arabic speech. This study focused on the problems associated with the orthography and spelling of dialects. The authors proposed an LF-MMI modeling framework for building the system. The system was trained using three LSTM, BLSTM, and TDNN techniques, separately. In LM training, two types of n-gram were used; firstly, tri-gram was used to generate decoding lattices, then, 4-gram was used for rescoring the output of the first type based on the external LM data. Both language models were trained using RNN with MaxEnt connections (RNNME) with Kneser–Ney as the smoothing technique. A multi-reference word error rate (MR-WER) of 25.3% was reported in this work as an average MR-WER for the Egyptian dialect.

Najafian *et al.* [30] presented a study for investigating the performance of several spoken dialect identification techniques. Multi-lingual such as Arabic, English, Czech, Hungarian, and Russian languages were trained separately for enhancing the accuracy. This work used the multi-dialectal speech corpus such as Egyptian (EGY), Gulf (GLF), Levantine (LEV), MSA, and North African. A new n-gram phonotactic feature was proposed and integrated with the SVMs classifier for generating the phone sequences. In addition, the i-vectors method was combined with the phonotactic features using DNN. Finally, convolutional neural networks (CNNs) were used to map the acoustic model and proposed features with each dialect language of the five dialects. The system achieved accuracies of 56.82% and 57.91% for phone n-gram with support vector machine (SVM) and phone n-gram with CNN, respectively.

Hassine *et al.* [31] built an Arabic ASR system for recognizing Arabic numbers (digits), from 0 to 9, in the Tunisian dialect. In the features extraction step, different techniques were used separately to extract features such as the perceptual linear prediction (PLP) technique, ΔPLP, MFCC, and vector quantization of Linde-Buzo-Gray (VQLBG). Then, all features were merged and used in training. The ANN type-known as feedforward back-propagation neural networks (FFBPNN)-was used for training the acoustic model. An average accuracy of 98.54% was obtained.

Khurana *et al.* [32] developed the DARTS system to convert speech to text in the Egyptian Arabic dialect. The transfer learning technique was used to transfer from the high-resource broadcast domain to the dialectal text. The acoustic model was trained on the Arabic MGB-2 and MGB-3 challenge using a deep neural network including a CNN and multiple layers of TDNN, and LSTM. Discriminative methods were used in training such as LF-MMI and Multi-LF-MMI. The training process was performed by the KALDI toolkit. Two LMs were developed: the first one was a tri-gram LM built by the SRILM toolkit with Kneser–Ney (KN) as the smoothing method. The other was 4-gram LM based on RNN-LM with MaxEnt connections using the Mikolov RNN LM toolkit. DARTS was evaluated using the MGB-3 testing corpus and achieved a WER of 35.8%.

Ali *et al.* [33] presented a new edition of the multi-genre broadcast challenge known as MGB-5. Its construction depends on the MGB-3 dataset and contains audio data collected from dialectal Moroccan recorded from over 48 hours from YouTube. These data were used for evaluating the Arabic ASR system. They proposed an LF-MMI modeling framework for building the system. The system was trained using three LSTM, BLSTM, and TDNN techniques, separately. RNNME was used to build 4-gram LM using the RNNLM toolkit. The features were transformed using adaptation techniques such as LDA, MLLT, and fMLLR. The KALDI toolkit was used for building all the trained models. Accuracies of 67.1% and 48.4% were obtained for AV-WER and MR-WER, respectively.

Ali *et al.* [34] evaluated the Arabic ASR system on the dialectal Arabic transcription that included a set of evaluation metrics. This work used these metrics by comparing their correlation with human judgments on a validation set of 1,000 utterances for six systems. They proposed a new degree for morphological abstraction and spelling normalization. The results showed that the new degree of morphological abstractions and spelling normalization demonstrated the best correlation with human judgment.

Bougrine *et al.* [35] developed a complete recipe for building large-scale speech corpus from web resources. The presented recipe was used to create a corpus for the Algerian Arabic dialect which was named KALAM'DZ. This corpus included eight classes of Algerian Arabic sub-dialects containing about 104.4 hours.

Alsharhan and Ramsay [36] evaluated the Arabic ASR system on the Arabic dialect based on MSA data. They used MFCCs for features extraction. The acoustic model was built based on the DNN network using the HTK toolkit. The pronunciation model was used for integrating the acoustic model with the LM depending on the pronunciation lexicon. The pronunciation lexicon contains a set of units (words) with single or multiple phonetic transcriptions. Then, LM is built using DNN and HMM. Two datasets were utilized for training and testing; the first dataset is the GALE phase 3 dataset for MSA, while the second is the Arabic dialect dataset which includes the Gulf, Iraqi, Egyptian, Levantine, and Maghrebi dialect version. The final integration achieved a WER between 3.24% and 5.35%.

Hamed *et al.* [37] collected and analyzed a speech corpus based on Egyptian and English conversations. Code-switching was used for mixing Arabic Egyptian and English conversations. Three-fold was proposed for building the corpus including conversational Egyptian Arabic spontaneous speech, obtaining manual transcriptions, and analyzing the speech from the code-switching perspective. Part-of-speech (POS) tags were used to annotate some of the transcriptions.

Ali [38] developed a multi-dialect ASR system for Arabic using an end-to-end approach. The author proposed CNN, RNN, and joint connectionist temporal classification (CTC)/attention encoder-decoder for building acoustic modeling. In LM training, an RNN with Kneser–Ney is used to build LM. An open-source corpus, collected from several corpora, was used for training and testing processes. An accuracy of 14.07% was obtained as WER.

Mubarak *et al.* [39] introduced the ASR system for dialectal Arabic speech using an end-to-end approach. They proposed a joint CTC/attention encoder-decoder for building acoustic modeling. In LM training, an RNN is used to build LM. The QASR corpus was used for training and testing processes. An average accuracy of 52.6% was reported in this work.

Hamed *et al.* [40] developed a system for switching Egyptian Arabic-English based on ASR. They used DNN-based hybrid and transformer-based depending on the end-to-end approach to build ASR systems. In LM training, an RNN is used to build LM. The MBG-3 corpus was used for training and testing processes. An optimal accuracy of 32.1% was obtained as WER.

Ahmed *et al.* [41] developed and described an Arabic ASR based on MGB-5 in Arabic. They applied speech augmentation using speed and volume perturbation, data reverberation, and music-noise-speech injection transformation. CNN with TDNN and TDNN-f were used for building the acoustic model. The x-vector and i-vector were combined and used as new features in this system. In addition, language model interpolation, semi-supervised learning, genre adaptation, and lattice-based MBR were proposed and combined. The proposed system achieved an average WER of 62.17%.

Al-Anzi and AbuZeina [42] presented dialectal Arabic speech system. This system includes pronunciation dictionaries, language models, and acoustic models. Acoustic model is trained and built based on a hybrid architecture Deep Neural Network Hidden Markov model (DNN-HMM) using HTK toolkit. N-gram language model is presented using ong-distance word relationships. MFCC is used to extract the features. The models are trained and evaluated by discrete-word speech dataset. The system achieved 54.02% as WER.

Hussein *et al.* [43] proposed an state-of-the-art end-to-end ASR for Arabic speech. They used transformer technqniue to build encoder and decoder. The language model is build using TDNN-LSTM. Mel filter bank is utilized to build acoustic feature. MGB3 and MGB5 corpora are used to train and evaluat system. The system achieved a new state-of-the-art performance at 27.5% and 33.8% for MGB3 and MGB5 respectively.

IV. DIALECTAL ARABIC SPEECH RECOGNITION SYSTEM

As mentioned in the above-mentioned literature review, most ASR systems comprise six steps: (i)

1) Feature extraction.
2) Lexical modeling.
3) Language modeling.
4) Acoustic modeling.
5) Discriminative criteria.
6) Evaluation.

The ASR system architecture is shown in Figure 2 and is further analyzed in this section.

A. FEATURE EXTRACTION

Feature Extraction is an important step of ASR tasks. The wave is formed in continuous size and time. The purpose of signal processing is to convert the waveform into vectors. Feature extraction is a process that is utilized to map the audio signals to a set of acoustic features that are utilized to build the acoustic model. The acoustic feature must be built without losing substantial signal data, minimizing variability across speakers, and environmental acoustic conditions, simultaneously. Moreover, it is used to distinguish speech from others. In addition, it utilizes extraction of the testing features as the input of the recognizer to generate the sequence of uttered words [44]. Thus, there are several techniques that can be used for feature extraction [45], [46].

- **Mel frequency cepstral coefficient (MFCC)** is a popular technique that is used to extract features of ASR. It depends on cepstral analysis, which is a method for separating speech signals into components in order to represent pitch and vocal tract information. MFCC simulates human behavior by distinguishing the sound frequencies since the frequency bands are calculated logarithmically. The feature processing begins with the windows step, which converts the waveform into vectors or chunks; optimally 25 ms are handled with 10 ms intervals. Then, each window is transformed into the spectral domain and power spectra using the short-time fast Fourier transform technique. In addition, power spectra are smoothed for each window using a 20-40 Mel filter bank. This smoothing method is utilized to calculate the frequency sensitivity of human hearing. The smoothed power spectra are logarithmically calculated in order to represent the Mel-filter bank (FBANK) features—that will be used in this work for training the acoustic models. The FBANK features will be prepared and presented for decorrelating discrete cosine transform (DCT) to produce MFCC [47]. This method has been used for feature extraction for different ASR systems (see [17–23]).
- **Perceptually based linear predictive analysis (PLP)** uses certain aspects of audition which provides the same spectral estimation of speech as LPC analysis but with a lower order model. In addition, it provides better performance for crossing speaker ASR. Furthermore, it is utilized to calculate the filter-bank filters followed by a linear predictive analysis and produce a cepstral representation. This method has been used for feature extraction for different ASR systems (see [9], [16], [17], [26]). LDA transformation is used to improve the separability and decrease data dimension in acoustic features.

Table 1 Studies that use these methods for feature extraction.

B. LEXICAL MODELING

A lexical model is a method for representing the phonemes sequence in the vocabulary. It is used as a pronunciation dictionary to map sequences of phones into words. Each line in the lexical model is utilized to represent a suitable word for the recognition model in the speech decoder with context-independent phonemes for these words. In addition, this lexical is a simple method for building lexical models. There are statistical methods to model lexical depending on the probabilities of multiple pronunciations of each word [48].

C. LANGUAGE MODELS

LM is statistical modeling (known as a model used in ASR decoding) for enhancing the word (unit) recognition process. LM depends on a large set of vocabularies that are each connected as sentences. In addition, LM is stored in a file that contains all words and their probability occurrences. The probability is the prior probability of a sequence of words and appears in the language. An ASR with LM is faster and achieves higher accuracy. In general, the quality of LM depends on the morphologically of language, i.e., LM of morphologically simplex language is better than LM of morphologically complex language. Thus, the LM of the Arabic language presents challenges due to the morphological complexity compared to some other languages [7], [49]. Sentences were not included in the ASR dataset and have an output model with zero probability—this is taken as a challenge and problem for the language model. To solve the zero-probability problem, the presence of a smoothing method endeavors to distribute probabilities to the sentences that have zero-based probability depending on the sentences in the dataset. Moreover, the method also tries to enhance the accuracy of the network model. There are a set of smoothing techniques that are used to calculate the probability of a word [49] and that are classified into backoff and interpolated techniques. In the first technique, the probability of the missed sentence in the corpus is estimated using its lower n-grams, while the second technique combines the sentence probability with its lower order, i.e., the probability of trigram, bigram, and unigram are combined. The smoothing techniques are Witten–Bell, Good–Turing, and Kneser–Ney smoothing [49]. Table 1 shows different techniques for building LM in different ASR systems. As shown, 25 dialectal Arabic ASR studies used LM.

D. ACOUSTIC MODELING

A set of statistical models are estimated to represent a set of sub-word/word (units), such as phonemes, tri-phones, or complete words. These models are usually used to measure how likely the acoustic features are emitted by the word

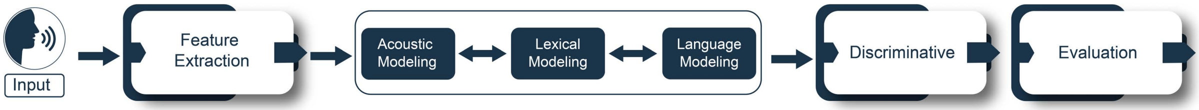


FIGURE 2. ASR system architecture.

TABLE 1. Summary of different studies based on various parameters (feature extraction, LM methods, AC methods, and adaption methods).

Study	Feature Extraction	LM Methods	AC Methods	Adaption Methods
[9]	PLP	NN with Kneser–Ney	HMM	FMLLR, MLLR, fMPE
[10]	MFFC	FST	HMM-based	MLLR, MAP, VTLN
[11]	MFCC	Kneser–Ney	HMM-based	
[12]	MFCC		HMM	
[13]	MFCC	CMU	HMM, GMM	MLLR, MAP
[14]	PLP, MFCCs, NN	NN with Kneser–Ney	SGMM	fMLLR, MLLR
[15]	MFCC	Kneser–Ney	GMM-HMM	
[16]	PLP		Cross-dialectal GMM	
[17]	PLP	Backoff LMs with entropy pruning, Katz	HMM	bMMI, LDA, CMLLR
[18]	MFCC	Standard types of LM	HMM	
[19]	MFCC	RNN-LM with Kneser–Ney	Ruled-based	
[20]	MFCC	RNN-LM with Kneser–Ney	GMM, SGMM, DNN	bMMI, MPE
[21]	MFCC	Kneser–Ney	GMM-HMM	MLLR, fMLLR, SAT, LDA
[22]	MFCC	RNNLM with Kneser–Ney	GMM, SGMM, DNN	bMMI, MPE
[23]	MFCC	RNNLM with Kneser–Ney	GMM, SGMM, Sequential DNN	bMMI, MPE
[24]	MFCC	RNNME	LSTM, BLSTM, TDNN	LDA, MLLT, fMLLR
[25]	MFCC	Neural network	GMM, HMM	
[26]	PLP		Ruled-based	
[27]	MFCC	Classical n-gram	DNN-HMM	sMBR
[28]	MFCC	RNNLM with Kneser–Ney	LSTM, BLSTM, TDNN	LDA, MPE, bMMI, fMLLR
[29]	MFCC	RNNME with Kneser–Ney	LSTM, BLSTM, TDNN	
[30]	MFCC, Bottleneck		CNN, SVM	LDA
[31]	PLP+Δ, MFCC+Δ, VQLBG		ANN	
[32]	MFCC	RNN-LM with Kneser–Ney	CNN, TDNN, LSTM	LF-MMI, Multi-LF-MMI
[33]	MFCC	RNNME	LSTM, BLSTM, TDNN	LDA, MLLT, fMLLR
[36]	MFCC	DNN, HMM	DNN	
[38]	MFCC	RNN with Kneser–Ney	CNN, RNN	CTC
[39]	MFCC	RNN	LSTM, BLSTM	CTC, Attention
[40]	MFCC	RNNLM	DNN-based, Transformer-based	CTC, Attention
[41]	MFCC with x-vector and i-vector	RNN-LM with Kneser–Ney	CNN with TDNN	Lattice-based MBR
[42]	MFCC	N-gram	DNN-HMM	
[43]	Mel filter bank	TDNN-LSTM	transformer	CTC

sequence hypothesis and its constituting sub-word units. The acoustic model is trained and built using generative learning algorithms. This model can recognize dialectal Arabic speech. In the reviewed studies, many techniques are used to represent acoustic modeling as shown in Table 1. A brief description of some of these techniques is presented.

1) HIDDEN MARKOV MODELS

HMMs were introduced at the end of the last century. HMMs are a special case of regular Markov models that have been evaluated as a powerful model for representing the time-varying signals as a parametric random process [50]. HMMs are considered the most popular acoustic models for ASR [6]. In addition, they are encoded by a finite state of the Markov model and decoded by a set of output distributions. The inputs of HMM are the temporal variability, while the outputs are spectral variability. The state of HMM is associated with probability density functions. The GMM with mixture diagonal covariance Gaussians are used to model each state in HMM [45], [45], [50]. Several diagonal covariance

Gaussians are utilized to generate probability densities as follows [50]:

$$b_i(x) = \sum_{j=1}^{N-i} w_{ij} N(x, \mu_{ij} \sigma_{ij}) \tag{1}$$

where *j* represents ranges over the count of Gaussian densities in the mixture of state S_i. The data likelihood will be maximized for training HMMs as follows [50]:

$$F_Y = \sum_{n=1}^{N} log(p(Y^n_{1:T_n} | w^n_{1:L_n})) \tag{2}$$

The parameters of the model (state transition probabilities and output distribution parameters, e.g., means and variances of a Gaussian) are automatically estimated from the training data. HMM is used to model the unit (phone, phoneme, word, etc.) [50]. The phone model is represented by a phoneme connected with each HMM. HMMs are used to model the phones as the main unit in speech, the left-to-right HMMs model each phone using three states with the input and stacked states. As shown in Table 1, nine dialectal Arabic ASR studies used HMM.

2) GAUSSIAN MIXTURE MODEL

GMM is a statistical learning model. GMM is a probabilistic model that represents the speech signals feature. It is used to manipulate variations in signals and convert them into a dynamic sequence of vectors. GMM is a suitable method that deals with text-independent ASR systems. To implement the likelihood ratio as a recognition model, the actual likelihood function must be determined. This function is selected based on the features that are extracted from signals. In addition, the GMM model is built depending on the underlying distribution of acoustic observations from speech. The temporal aspects of the utterance do not impact GMM modeling. In speech recognition, each utterance is represented as a GMM for producing this model. The parameters in model λ must be estimated in order to match the best training vectors in utterance [51]. The most favorable techniques for estimating these parameters in the model are:

1) Maximum likelihood (ML) estimation. The main goal of ML is to obtain the model parameters that maximize the likelihood of the GMM;
2) Expectation-maximization (EM). EM is an iterative algorithm utilized to estimate ML of GMM parameters.

As shown in Table 1, seven dialectal Arabic ASR studies used GMM.

3) SUBSPACE GAUSSIAN MIXTURE MODEL

In conventional acoustic models, a GMM with a large set of parameters can represent every HMM distribution. The SGMM also represents the states' distribution with a small set of parameters as low dimensional subspace. There is a high correlation between states' distributions, therefore, the distributions of states can be represented by a low dimensional subspace for all states. Human sounds correspond to a limited variety of distributions, therefore, speech is considered as triphone states with a high correlation between their distributions. The SGMM is suitable for ASR and comprises shared parameters for all states. In addition, SGMMs can be naturally trained in a multilingual fashion. However, in an SGMM, the correlations across the triphone states are stored in a low dimensional model subspace as parameters [52]. All context-dependent HMM states in SGMMs use the universal background model (UBM) for sharing a common representation. UBM represents a GMM model trained over whole speech classes that are pooled together [53]. A GMM-UBM is a large mixture of Gaussians that represents all speech with I components; it is used for pruning the Gaussian components and initializing the model. The acoustic space is split into I regions by UBM, where the acoustics are defined using M_j, N_i, and w_i. In UBM, the selected highest P Gaussian components with maximum likelihood scores are used in both model training and recognition. As in GMM, we used ML to estimate the SGMM parameters, and EM to estimate ML of SGMM. As shown in Table 1, four dialectal Arabic ASR studies used SGMM.

4) DEEP NEURAL NETWORK MODEL

A DNN model comprises an input layer, an output layer, and two or more layers of hidden units. Each hidden unit is used to associate all inputs from the previous layer to the scalar state using the logistic function and sent into the next layer. DNNs are discriminatively trained using the backpropagation of cost function derivatives. This backpropagation is utilized to determine the conflict between the original outputs and resulted from DNN training [54]. In the softmax function, the cross-entropy between the probabilities d and softmax output p is represented by the natural cost function C as follows:

$$C = -\sum_{j} d_j log(p_j) \tag{3}$$

where the probability *d* is one or zero.

DNN can be trained on a large training set by calculating the derivatives on a small part of the training set "minibatch" compared to the whole training set. Then, it updates the weights to the gradient. The trained neural networks in DNN are used to recognize speech. It includes the dimension of the input spectral features as the input layer, N hidden layers, and one output layer. The output layer dimension is equal to the number of utterances the system is designed to identify. The frame-level DNN posteriors from the output layer must be combined by simply averaging over the test utterance [55]. DNN can be used to train and recognize speech signals during a low resource system without a secondary classifier. The secondary classifier is unsuitable for small datasets, which requires computational resources. Increasing the number of hidden layers will enhance the system's performance, however, the complexity will be increased. As shown in Table 1, 12 dialectal Arabic ASR studies used DNN.

5) CONVOLUTIONAL NEURAL NETWORK

CNN [56] is a kind of DNN. It has a mechanism for simulating the mammal visual neuron systems [57] which activate neurons in specific areas in the visual field. CNN has conditioned—as opposed to fully integrated—connections to manipulate data using a grid-form essential structure. For example, an image can be represented by 2D pixel grids and fixed-length audio can be represented by 1D grids. In addition, CNN has novel properties that render DNN more suitable for image and signal data. CNN has three stages: a convolution, detector, and pooling stage.

- **Convolution**: The convolution stage is considered the main component of CNN models. Practically, each value in the feature vector is connected with the nearest values in the feature vectors. Thus, favorable features are calculated locally and selected. The convolution includes a process to handle noise in local regions [58]. Overall, this process trains the CNN model well.
- **Detector**: The detector stage receives the convolution outputs and applies the nonlinear activation functions to generate high-level features.
- **Detector**: A final pooling stage is utilized to adapt the activation function resulting from the decoder stage. This stage integrates outputs and shows signal data for different local regions. Pooling is more suitable to represent small fluctuations that are obtained from inputs. In addition, pooling is used to decrease outputs compared to the detector stage, thus, reducing time and computational complexities.

Because a CNN is a simulated biologically inspired model, it is, therefore, suitable to develop acoustic models in ASR systems in order to enhance the performance. In addition, the structural locality from the acoustic feature is used to reduce the spectral variance in acoustic features and long-term dependencies in the speech frames by taking prior speech signal knowledge [59], [60]. Sainath *et al.* [61] reported that CNN achieves 13–30% enhancement over GMMs, and 4–12% enhancement over DNNs, using 700 hours of speech data. As shown in Table 1, four dialectal Arabic ASR studies used CNN.

6) TIME-DELAY NEURAL NETWORK

Time-delay neural networks (TDNNs) are types of CNNs that are used for sharing the weights in a single temporal dimension. The first TDNN model was proposed to recognize phonemes [62]. Then, TDNNs were utilized for recognizing the spoken word [63] and handwriting [64], enabling the acoustic model to learn the temporal dynamics of the speech signal using short-term acoustic feature vectors. Moreover, it uses sub-sampling for reducing computation in training. In DNN, the wider temporal context is processed by a wide contextual window of features in the initial layer, while in TDNN, each layer corresponds to a different level of the entire features—local patterns in the entire features are learned by the first layer and higher layers are used to learn a wider temporal context. Each layer in a TDNN is operated at a different temporal resolution, which is increased as one moves deeper into the network. As shown in Table 1, six dialectal Arabic ASR studies used TDNN.

7) LONG SHORT-TERM MEMORY NETWORK

LSTMs are a special type of RNNs used for the evolution of RNN. The LSTM method can save information over a long period using long-term dependencies in order to find and exploit long-range context. The standard RNN has a single neural network, while LSTM uses four interacting layers with a unique communication link [4], [65]. In ASR, we can use the coming context as well if the transcription for all utterances is obtained at training time. An LSTM calculates an input sequence $X = x_1, x_3, ..., x_T$ and the corresponding output sequence $Y = y_1, y_2, ..., y_L$ using the calculation of the network unit activation. An LSTM is used in the training stage with sub-sampling given the T-length of the speech feature sequence o_{t-1}. It is utilized to produce a high-level feature $h_{1:T_0}$ as follows:

$$h_t = LSTM(x_t, o_{t-1}) \tag{4}$$

where h denotes the sub-sampling. The input features X will be handled to create the hidden states h_t based on frame-wise operations. LSTM presents the outputs to reduce the computational cost. Therefore, in ASR, the input length is different from the output length [66]. A bidirectional LSTM (BLSTM) is an LSTM in the hidden layers. As shown in Table 1, 11 dialectal Arabic ASR studies used LSTM and BLSTM.

E. TRANSFORMER MODEL

The Transformer model architecture is the same as sequence-to-sequence attention-based models except relying entirely on self-attention and position-wise, fully connected layers for both the encoder and decoder [67], [68]. The transformer-based model comprises two parts: an encoder with a set of blocks; a decoder with a set of blocks [69]. The encoder maps an input sequence of symbol representations $x = (x_1, \ldots, x_n)$ to a sequence of continuous representations $z = (z_1, \ldots, z_n)$. Given z, the decoder then generates an output sequence $y = (y1, \ldots, y_m)$ of symbols one element at a time. Transformer learns sequential information via a self-attention mechanism instead of the recurrent connection employed in RNN [70].

An attention function maps a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The subsampled sequences (X-sub), that are generated by the previous step, represent the input to the encoder blocks. Encoder transform (X-sub) to (Q, K, V) using a self-attention layer with a Softmax as follows [67], [68]:

$$Self\ Attention(Q, K, V) = softmax(\frac{Q * K^T}{\sqrt{d^k}}) * V \tag{5}$$

$Q \in \mathbb{R}^{n^q * d^q}$, $K \in \mathbb{R}^{n^k * d^k}$, and $V \in \mathbb{R}^{n^v * d^v}$ denote to queries, keys, and values respectively. $d*$ is the dimensions of values, keys, and queries, and $n*$ is sequence lengths.

The multi-head attention (MHA) to perform multiple attention networks. MHA yielded from all concatenated self-attention heads as follows:

$$MHA(Q, K, V) = [H_1, H_2, \ldots, H_h]W^h \tag{6}$$
$$H_i = Self\ Attention(Q_i, K_i, V_i) \tag{7}$$

where h denotes the attention heads number in a single layer and i is the i^th head in the layer. The MHA output is normalized before being sent into the Feed Forward (FF) sub-layer linked network, which is implemented for every point individually as follows.

$$FF(h[t]) = max(0, h[t] * W_1 + b_1)W_2 + b_2 \tag{8}$$

where h[t] denotes the t^th position of the input H to the FF sub-layer.

F. DISCRIMINATIVE CRITERIA

In speech recognition, the acoustic model may be trained using large datasets consisting of hundreds of hours (or greater), from different speakers. However, there are often utterances that are poorly represented in the training data. This leads to a conflict in mapping between training and testing representation. To solve this problem and reduce the mismatch, there are adaptation techniques for discriminative training of acoustic models. The discriminative training approach directly optimizes a mapping function from the input samples to output labels and is used to enhance the acoustic model for recognizing utterances [71], [72]. Therefore, the main goal of the discriminative learning approach is to modify only the decision boundary without constructing a data generator from the entire feature space. As in the reviewed studies, several discriminative training criteria are used for the dialectal Arabic speech recognition, such as LDA, MMLR, fMMLR, maximum mutual information estimation (MMIE), boosted MMI, MPE, CTC, and attention-based models (see Table 1). In addition, adaptation is an optimal approach that alleviates conflicts between the models and the data from any utterance, channel, or another factor. As shown in Table 1, 19 dialectal Arabic ASR studies used adaptation methods.

G. EVALUATION

As in the reviewed studies, accuracy performance metric is used for evaluating the performance of ASR systems. Moreover, the perplexity metric is used for evaluating the performance of LM. This section describes these evaluation metrics in some detail.

1) ASR EVALUATION

The performance evaluation of ASR is usually presented in terms of two criteria: (1) accuracy (Acc), which represents the percentage of the accuracy, and (2) WER, which represents the percentage of the word-level errors of the recognized units. These criteria are defined as follows:

$$SER = \frac{\#correctly\_recognized\_sentences}{\#All\_Sentences} * 100 \tag{9}$$

$$WER = \frac{S + D + I}{N} * 100 \tag{10}$$
$$Acc = 100 - WER \tag{11}$$

where N represents all the words in the set of evaluation utterances, substitutions (S) denotes the number of misrecognized words, deletions (D) represents the number of the deleted words in the recognition result, and I is the number of the inserted words in the recognition result.

2) LANGUAGE MODEL EVALUATION

The performance evaluation of LM uses the perplexity measure, which is dependent on the token in transcriptions. The perplexity of LM is calculated for K tokens as follows [73]:

$$Perplexity = (\prod_{i=1}^{K} P(token_i|token_{j<i}))^{-\frac{1}{K}} \tag{12}$$

where $P(token_i|token_{j<i})$ represents the probability of i-th throw of the LM training depending on the first $i - 1$ tokens.

Table 2 Shows the datasets and evaluation for all systems that are presented in the reviewed studies.

V. DISCUSSION AND CHALLENGES

In this section, we present a discussion and analysis of the results and highlight some challenges for the dialectal Arabic ASR system.

A. DISCUSSION AND ANALYSIS

As in the literature review, many studies present the dialectal Arabic ASR using several techniques and methods as shown in Table 1. From Table 1, we can see that most studies used the MFCC technique for feature extraction. These studies used 13 MFCC with 39 and 40-dimensional high-resolution Mel frequency cepstral coefficients. Furthermore, some of these studies used LDA for features transformation, while other studies used techniques such as PLP (12- and 13-dimensional PLP), neural network (NN), bottleneck, and VQLBG. In addition, most studies used language modeling using different techniques in which a large number of these studies used RNN for building language models. Moreover, some of these studies used the Kneser–Ney smoothing technique for enhancing LM. The 3-gram and 4-gram LMs were used in these studies. Five studies did not use LM. Table 1 shows the techniques and approaches that were used for building acoustic models in the presented studies. Most studies used traditional techniques and others used deep learning techniques for building acoustic models. Moreover, some studies used hybrid techniques. Furthermore, two studies used the ruled-based technique for building models. Most studies used adaptation (discriminative) techniques, as shown in Table 1. These techniques were used to enhance the acoustic model for recognizing utterances. In general, most studies used a traditional approach for building the dialectal Arabic ASR system, while two studies used an end-to-end approach.

We cannot decide the state-of-the-art of dialectal Arabic ASR systems. However, according to our literature review

TABLE 2. Comparison of different studies based on parameters (dataset, ACC(%), WER(%), perplexity, and year).

Study	Dataset	ACC(%)	WER(%)	Perplexity	Year
[9]	LDC GALE Phase 3		24.7		2009
[10]	ECA	99.34			2009
[11]	LDC Iraqi-Arabic		33.3		2009
[12]	ALGASD	91.65			2010
[13]	ECA		35	19.60%	2010
[14]	FBIS and TDT-4		9.1	9.10%	2010
[15]	ECA		13.4		2011
[16]	Babylon Levantine Arabic		39		2012
[17]	Collected dataset		24.8		2012
[18]	Multi-dialect Arabic parallel speech corpus		14.6		2013
[19]	TARIC		9		2014
[20]	Collected dataset		44.71	2047	2014
[21]	Qatari Arabic corpus		64.4	315.5	2014
[22]	Collected dataset		23	2047	2015
[23]	Collected dataset	80.9		2020	2015
[24]	MGB-2		14.2	400	2016
[25]	MCSM database		57.9		2017
[26]	Tunisian dialect		22.6		2018
[27]	Nemlar, Gigaword Arabic, and NetDC		89		2017
[28]	MGB-3		37.5		2017
[29]	Collected dataset		25.3		2017
[35]	Collected corpus				2017
[30]	Multi-dialectal speech corpus	56.82			2018
[31]	Arabic Tunisian digit	98.54			2018
[37]	Collected corpus	79.8			2018
[32]	MGB-2 and MGB-3		35.8	481	2019
[33]	MGB-5	67.1			2019
[34]	Collected dataset		41.67		2019
[41]	MGB-5		62.17		2019
[36]	GALE phase 3 and collected dataset		23.1		2020
[38]	Open-source corpus		14.07		2020
[39]	QASR corpus	52.6			2021
[40]	CALLHOME, MGB-3 corpus		32.1	566	2022
[42]	Discrete-word speech		54.02		2022
[43]	MGB3 and MGB5		27.5 and 33.8		2022

TABLE 3. Number of studies over the last five years.

Year	No. of Studies
2017	6
2018	3
2019	4
2020	2
2021	1
2022	3
Total studies	**19**

and current knowledge, we can conclude that studies implementing the end-to-end approach are at the front of dialectal Arabic ASR systems as shown in Table 2.

From Table 2, we can observe that the WER term was reported as result for most studies, while other studies used the accuracy terms. In addition, most studies did not report the perplexity of LM. Furthermore, we can see that five studies are presented in 2017 and one study is presented in 2016. Over the last five years, 17 studies were presented as shown in Table 3.

As shown from the literature review and Table 2, datasets, corpora, and databases were used in the dialectal Arabic ASR systems. The corpora and datasets were collected and built as follows:

Audio speech files are collected from radio and/or TV broadcast news, telephone conversation, and YouTube channels based on the same conditions. Audio files will be separated into smaller length. Then, the audio files will be converted into wav format. After that, speech files will be re-sampled into the same sampling rate. The corresponding transcript file for each audio file will be created. Finally, the transcript will be converted into Buckwalter format.

Some of these data sources are freely available and others require an access fee. Table 4 summarizes the characteristics and availability of some (free) data sources that were used in dialectal Arabic ASR systems.

The Arabic dialect has several types of dialects such as Algeria (DZ), Egypt (EG), Iraq (IQ), Jordan (JO), Saudi Arabia (SA), Kuwait (KW), Lebanon (LB), Libya (LY), Mauritania (MR), Morocco (MA), Oman (OM),

TABLE 4. Characteristics of data source.

Data Source	Characteristics	Free
GALE Phase 2	GALE Phase 2 Arabic Broadcast Conversation was collected by linguistic data consortium (LDC). It includes various Arabic dialects from different Arabic regions and countries regional Arabic such as Gulf region, Levantine countries, and Egypt. It was collected from Arabic programming from various sources (TVs and channels). It contains about 123 hours of Arabic conversation speech. This dataset contains 143 audio files presented in flac and sampled in 16 kHz and 16 bit.	No
GALE Phase 3	GALE Phase 3 Arabic Broadcast Conversation was created based on GALE Phase 2 and collected by LDC. It includes various dialects of Arabic from different Arabic regions and countries regional Arabic such as the Gulf region, Levantine countries, and Egypt. It was collected from Arabic programming from various sources (TVs and channels). It contains about 132 hours of Arabic conversation speech. This dataset contains 175 audio files presented in flac and sampled in 16 kHz and 16 bit.	No
ECA	ECA was recorded by Hama PC headset CS-499. It includes the speech data of Egyptian dialectal. This corpus consists of 50 Egyptian native speakers, 50% for male speakers, and 50% for female speakers. The age of the speakers is between 18 and 32 years old. The total utterances of this corpus are 2500, 50 utterances for each speaker. Utterances represent several domains such as greetings, time and dates, words spelling, restaurants, train reservation, Egyptian proverbs. It was sampled at 16 kHz and 16 bit.	No
Iraqi-Arabic corpus	This dataset is Iraqi Arabic Conversational Telephone Speech. It comprises 3k minutes of dialectal Iraqi speech. It includes 478 conversations that are recorded by 474 unique speakers. The average duration of the recorded file is about 6 minutes.	No
ALGASD	ALGASD corpus represents the dialectal Algerian speakers from different regions of Algeria. It includes speech data recording by 300 speakers from 11 regions. The age of the speakers is between 18 and 50 years old.	No
Babylon Levantine Arabic speech corpus	The Babylon Levantine Arabic Speech corpus represents the speech data of Levantine dialect Arabic. It consists of 164 speakers, 101 males, and 63 females. It contains about 45 hours segmented into 79500 audio.	No
Multi-dialect Arabic parallel speech corpus	The multi-dialect Arabic parallel speech corpus includes MSA, Gulf, Levantine, and Egyptian dialects. It contains about 32 hours segmented into 67000 audio. The speech data in this corpus were collected from the travel and tourism domain.	No
TARIC	TARIC includes the speech data of the Tunisian Arabic dialect. It contains 20 hours, 4662 dialogues, 18657 statements, and 71684 words.	No
Qatari Arabic Corpus	Qatari Arabic Corpus includes the speech data of Qatari Arabic dialect. It was collected from TV series and talk show programs. It contains 15 hours and is sampled at 16 kHz, and 16 bits.	Yes
MGB-2	MGB-2 includes the speech data of MSA and various dialectal Arabic such as Egyptian, Gulf, Levantine, and North African. The total duration of this dataset is 1.2k hours. MSA represents 70% of speech data, while dialectal Arabic represents other speech data.	Yes
MCSM database	MCSM is the Maghrebian broadcasts from Algeria, Morocco, and Tunisia. It contains 53 hours. It was collected from TV media including entertainment and talk show programs.	No
FACST	The FACST corpus includes the speech data of the Maghrebian dialectal Arabic. It contains eight speakers; the age of these speakers is between 20 and 35 years old.	No
NEMLAR	The Nemlar is a Broadcast News Speech Corpus for Algerian dialectal Arabic. It was collected from different radio stations. It consists of about 40 hours of recording by different speakers.	No
NetDC	The NetDC is a Broadcast News Speech Corpus for Algerian dialectal Arabic. It was collected from one radio station. It consists of about 22.5 hours of recording by different speakers.	No
MGB-3	MGB-3 includes the speech data of Egyptian dialectal Arabic. It was collected from YouTube recordings. It comprises seven different genres comedy, cooking, family/kids, fashion, drama, sports, and science. It contains 16 hours extracted from eighty YouTube videos. It is inadequate by itself for developing accurate ASR systems.	Yes
Multi-dialectal speech corpus	Multi-dialectal speech corpus includes MSA, Egyptian, Gulf, Levantine, and North African. It comprises different speakers. It contains 120.2 hours, 17016 sentences, and 405k words.	No
Arabic Tunisian digit	Arabic Tunisian digit includes the speech data of Tunisian digit. It includes ten speakers (5 males and 5 females); the ages of these speakers are between 9 and 60 years old. It was recorded in Mono wave files and sampled at 44kHz and 16 bits.	No
MGB-5	MGB-5 includes the speech data of Moroccan dialect Arabic. It was collected from YouTube recordings. It was collected from YouTube recordings. It comprises 7 different genres comedy, cooking, family/kids, fashion, drama, sports, and science. It contains 48 hours extracted from 93 programs on YouTube.	Yes
Open-source corpus	The open-source corpus covers MSA and multiple Arabic dialects. It comprises speech data recorded by 450–500 Arabic speakers (males and females) with different dialects. The ages of speakers are between 35 and 60 years old. It is collected from other corpora such as KACST, KSU, MGB-2, etc. It contains 1462 items over about 400 hours for dialectal Arabic.	Yes
QASR corpus	QASR corpus covers MSA and multiple Arabic dialects. It comprises MGB-2 and newly collected speech data. It contains 2000 hours.	Yes
CALLHOME	The CALLHOME Egyptian Arabic corpus is the speech data of telephone conversations. It contains 120 telephone conversations. It includes many male and female speakers.	No
BOLT Egyptian Arabic treebank	BOLT Egyptian Arabic Treebank corpus is the speech data of telephone conversations. It was collected by LDC. It consists of 153,171 tokens and 182,965 split tokens.	No

Palestine (PS), Qatar (QA), Sudan (SD), Syrian Arab Republic (SY), United Arab Emirates (AE), Yemen (YE), Tunisia (TN), and Bahrain (BH). Figure 3 shows the number of speakers from each dialect type.

TABLE 5. Summary of the Arabic dialect types.

Dialect Studies	DZ	EG	IQ	JO	SA	KW	LB	LY	MR	MA	OM	PS	QA	SD	SY	AE	YE	TN	BH
[10]		√																	
[11]			√																
[12]	√																		
[13]		√																	
[15]		√																	
[16]				√			√					√			√				
[17]		√		√	√		√									√			
[18]		√		√	√	√	√				√	√	√		√	√			√
[19]																		√	
[20]		√																	
[21]													√						
[22]	√	√						√	√	√								√	
[23]		√																	
[25]	√									√								√	
[26]																		√	
[27]	√																		
[28]		√																	
[29]		√																	
[30]	√	√	√	√	√	√		√		√	√	√			√	√		√	√
[31]																		√	
[32]		√																	
[33]										√									
[35]	√																		
[36]	√	√	√	√	√	√	√			√	√	√	√		√	√		√	√
[37]		√																	
[39]													√						
[40]		√																	
[41]										√									
[42]		√																	
[43]		√																√	
No. of Studies	**7**	**17**	**3**	**5**	**5**	**3**	**4**	**2**	**1**	**6**	**3**	**4**	**4**	**0**	**4**	**4**	**0**	**8**	**3**

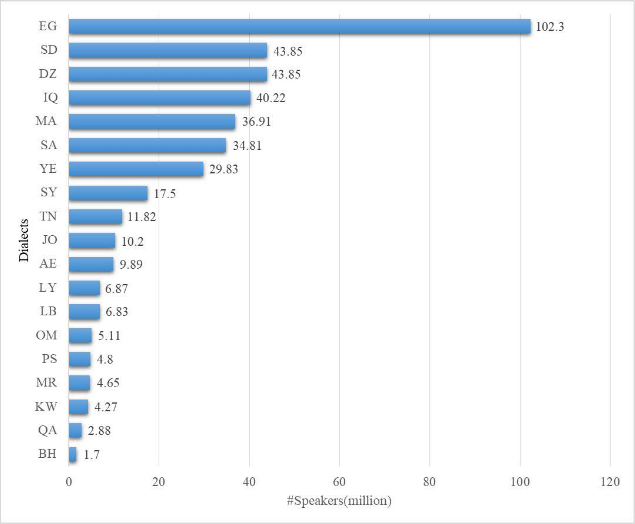


FIGURE 3. Number of speakers in each type of Arabic dialect.

The presented studies show different dialectal Arabic ASR systems for some of these Arabic dialects. Table 5 summarizes the types of Arabic dialects in the presented studies. From this table, we note that the Egyptian dialect has the highest study count (17), while Algeria and Tunisia have seven and eight studies, respectively. Mauritania-type dialect has only one study representing the lowest. In addition, Sudan and Yemeni dialects have no studies.

Some of the presented studies were published in the Web of Science and Scopus databases as shown in Table 6.

Figure 4 shows a comparison between the types of Arabic dialects regarding the number of studies for each type.

B. CHALLENGES

As mentioned above, many studies have developed dialectal Arabic ASR systems. Arabic dialects that includes vocalized components and using morphological decomposition to address the challenges of dealing with the huge lexical variety. However, adding to the challenges that are mentioned in the introduction, many challenges still exist:

1) There are a few counts of datasets that are related to the Arabic dialect. This leads to develop a bad ASR system.
2) There are no datasets for some dialect types such as Sudan and Yemeni. Therefore, most dialectal Arabic ASR systems can not recognize all Arabic dialect languages.

TABLE 6. Summary of studies published in web of science and scopus databases.

Studies	Web of Science	Scopus Database
[9]		√
[10]		√
[12]	√	
[14]		√
[15]		√
[16]	√	√
[17]		√
[26]	√	√
[27]		√
[30]		√
[36]	√	√
[38]		√
[40]	√	√
[42]	√	√
[43]	√	√
#Studies	**7**	**14**

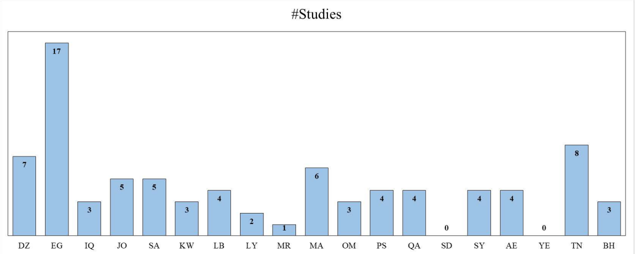


FIGURE 4. Comparison of Arabic dialect types regarding the number of studies.

3) All of the available studies used a non-diacritized Arabic version. The non-diacritized Arabic word may have many meaning with the word meaning to its' position in the word context. Thus, the accuracy of training acoustic model may be decreased. Also LM may be less predictive.
4) There are studies that lack variation in data.
5) All of the studies use English letters instead of Arabic letters in transcription using the Buckwalter format. Thus, the data needs more efforts.
6) Many of the presented studies used the MSA version to train and adapt the acoustic model, and dialectal version for testing; i.e., these studies did not use pure dialectal data in the training process.
7) The studies that were developed based on the collected or specific datasets have good accuracy. These studies did not use the standard datasets. Moreover, some of these studies used one dialect type.
8) The multi-dialect systems have a low accuracy compared to one-dialect systems.

VI. CONCLUSION

In this work, we reviewed 35 studies of the dialectal Arabic ASR. Many approaches and techniques were described in feature extraction, including lexical modeling, language modeling, acoustic modeling, discriminative criteria, and evaluation steps. Moreover, we presented the current progress of the dialectal Arabic ASR and introduced three comparisons between the presented studies including techniques and methods, datasets, accuracies, dialect types—these studies were analyzed and discussed. In addition, a brief of the dialectal Arabic data sets and corpora is presented. We also discussed and highlighted some challenges and problems. Due to challenges and from the analysis, we suggest some future studies including collecting diacritized data, collecting new various data, collecting Sudanese and Yemeni dialect data, adapting techniques and methods to address Arabic letters, and applying other techniques and methods for building the dialectal Arabic speech system.

REFERENCES

[1] X. He and L. Deng, "Discriminative learning for speech recognition: Theory and practice," *Synth. Lect. Speech Audio Process.*, vol. 4, no. 1, pp. 1–112, Jan. 2008.

[2] D. AbuZeina, W. Al-Khatib, M. Elshafei, and H. Al-Muhtaseb, "Cross-word Arabic pronunciation variation modeling for speech recognition," *Int. J. Speech Technol.*, vol. 14, no. 3, pp. 227–236, Sep. 2011.

[3] A. Boumehdi and A. Yousfi, "Arabic speech recognition independent of vocabulary for isolated words," in *Proc. 6th Int. Congr. Inf. Commun. Technol.* Singapore: Springer, 2022, pp. 585–595.

[4] A. A. Abdelhamid, H. A. Alsayadi, I. Hegazy, and Z. T. Fayed, "End-to-end Arabic speech recognition: A review," in *Proc. 19th Conf. Lang. Eng. (ESOLEC)*, 2020, pp. 233–249.

[5] P. Cardinal, A. Ali, N. Dehak, Y. Zhang, T. A. Hanai, Y. Zhang, J. R. Glass, and S. Vogel, "Recent advances in ASR applied to an Arabic transcription system for Al-Jazeera," in *Proc. 15th Annu. Conf. Int. speech Commun. Assoc.*, Sep. 2014, pp. 2088–2092.

[6] O. Hamed and T. Zesch, "A survey and comparative study of Arabic diacritization tools," *J. Lang. Technol. Comput. Linguistics*, vol. 32, no. 1, pp. 27–47, 2017.

[7] S. M. Abdou and A. M. Moussa, "Arabic speech recognition: Challenges and state of the art," in *Computational Linguistics, Speech and Image Processing For Arabic Language*. Singapore: World Scientific, 2019, pp. 1–27.

[8] M. Elmahdy, R. Gruhn, and W. Minker, *Novel Techniques for Dialectal Arabic Speech Recognition*. Springer, 2012.

[9] H. Soltau, G. Saon, B. Kingsbury, H.-K. J. Kuo, L. Mangu, D. Povey, and A. Emami, "Advances in Arabic speech transcription at IBM under the DARPA GALE program," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 5, pp. 884–894, Jul. 2009.

[10] M. Elmahdy, R. Gruhn, W. Minker, and S. Abdennadher, "Modern standard Arabic based multilingual approach for dialectal Arabic speech recognition," in *Proc. 8th Int. Symp. Natural Lang. Process.*, Oct. 2009, pp. 169–174.

[11] H. Al-Haj, R. Hsiao, I. Lane, A. W. Black, and A. Waibel, "Pronunciation modeling for dialectal Arabic speech recognition," in *Proc. IEEE Workshop Automat. Speech Recognit. Understand.*, Dec. 2009, pp. 525–528.

[12] S. A. Selouani and M. Boudraa, "Algerian Arabic speech database (ALGASD): Corpus design and automatic speech recognition application," *Arabian J. Sci. Eng.*, vol. 35, no. 2, pp. 157–166, 2010.

[13] M. Elmahdy, R. Gruhn, W. Minker, and S. Abdennadher, "Cross-lingual acoustic modeling for dialectal Arabic speech recognition," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2010, pp. 873–876.

[14] G. Saon, H. Soltau, U. Chaudhari, S. Chu, B. Kingsbury, H.-K. Kuo, L. Mangu, and D. Povey, "The IBM 2008 GALE Arabic speech transcription system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4378–4381.

[15] M. Elmahdy, R. Gruhn, S. Abdennadher, and W. Minker, "Rapid phonetic transcription using everyday life natural chat alphabet orthography for dialectal Arabic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 4936–4939.

[16] P.-S. Huang and M. Hasegawa-Johnson, "Cross-dialectal data transferring for Gaussian mixture model training in Arabic speech recognition," *Constraints*, vol. 1, p. 1, 2012.

[17] F. Biadsy, P. J. Moreno, and M. Jansche, "Google's cross-dialect Arabic voice search," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4441–4444.

[18] K. Almeman and M. Lee, "A comparison of Arabic speech recognition for multi-dialect vs. specific dialects," in *Proc. 7th Int. Conf. Speech Technol. Hum.-Comput. Dialogue (SpeD)*, 2013, pp. 16–19.

[19] A. Masmoudi, M. E. Khemekhem, Y. Esteve, L. H. Belguith, and N. Habash, "A corpus and phonetic dictionary for Tunisian Arabic speech recognition," in *Proc. 9th Int. Conf. Lang. Resour. Eval.*, 2014, pp. 306–310.

[20] A. Ali, H. Mubarak, and S. Vogel, "Advances in dialectal Arabic speech recognition: A study using Twitter to improve Egyptian ASR," in *Proc. Int. Workshop Spoken Lang. Transl. (IWSLT)*, 2014, pp. 156–162.

[21] M. Elmahdy, M. Hasegawa-Johnson, and E. Mustafawi, "Development of a TV broadcasts speech recognition system for Qatari Arabic," in *Proc. LREC*, 2014, pp. 3057–3061.

[22] S. Wray, H. Mubarak, and A. Ali, "Best practices for crowdsourcing dialectal Arabic speech transcription," in *Proc. 2nd Workshop Arabic Natural Lang. Process.*, 2015, pp. 99–107.

[23] A. Ali, W. Magdy, and S. Renals, "Multi-reference evaluation for dialectal speech recognition system: A study for Egyptian ASR," in *Proc. 2nd Workshop Arabic Natural Lang. Process.*, 2015, pp. 118–126.

[24] S. Khurana and A. Ali, "QCRI advanced transcription system (QATS) for the Arabic multi-dialect broadcast media recognition: MGB-2 challenge," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLTW)*, Dec. 2016, pp. 292–298.

[25] D. Amazouz, M. Adda-Decker, and L. Lamel, "Addressing code-switching in French/Algerian Arabic speech," in *Proc. INTERSPEECH*, Aug. 2017, pp. 62–66.

[26] A. Masmoudi, F. Bougares, M. Ellouze, Y. Estève, and L. Belguith, "Automatic speech recognition system for Tunisian dialect," *Lang. Resour. Eval.*, vol. 52, no. 1, pp. 249–267, Mar. 2018.

[27] M. A. Menacer, O. Mella, D. Fohr, D. Jouvet, D. Langlois, and K. Smaïli, "Development of the Arabic loria automatic speech recognition system (ALASR) and its evaluation for Algerian dialect," *Proc. Comput. Sci.*, vol. 117, pp. 81–88, Jan. 2017.

[28] A. Ali, S. Vogel, and S. Renals, "Speech recognition challenge in the wild: Arabic MGB-3," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 316–322.

[29] A. Ali, P. Nakov, P. Bell, and S. Renals, ''WERD: Using social text spelling variants for evaluating dialectal speech recognition,'' in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 141–148.

[30] M. Najafian, S. Khurana, S. Shan, A. Ali, and J. Glass, ''Exploiting convolutional neural networks for phonotactic based dialect identification,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5174–5178.

[31] M. Hassine, L. Boussaid, and H. Massaoud, ''Tunisian dialect recognition based on hybrid techniques,'' *Int. Arab J. Inf. Technol.*, vol. 15, no. 1, pp. 58–65, 2018.

[32] S. Khurana, A. Ali, and J. Glass, ''DARTS: Dialectal Arabic transcription system,'' 2019, *arXiv:1909.12163*.

[33] A. Ali, S. Shon, Y. Samih, H. Mubarak, A. Abdelali, J. Glass, S. Renals, and K. Choukri, ''The MGB-5 challenge: Recognition and dialect identification of dialectal Arabic speech,'' in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 1026–1033.

[34] A. Ali, S. Khalifa, and N. Habash, ''Towards variability resistant dialectal speech evaluation,'' in *Proc. INTERSPEECH*, Sep. 2019, pp. 336–340.

[35] S. Bougrine, A. Chorana, A. Lakhdari, and H. Cherroun, ''Toward a web-based speech corpus for Algerian dialectal Arabic varieties,'' in *Proc. 3rd Arabic Natural Lang. Process. Workshop*, 2017, pp. 138–146.

[36] E. Alsharhan and A. Ramsay, ''Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition,'' *Lang. Resour. Eval.*, vol. 54, no. 4, pp. 975–998, Dec. 2020.

[37] I. Hamed, M. Elmahdy, and S. Abdennadher, ''Collection and analysis of code-switch Egyptian Arabic–English speech corpus,'' in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2018.

[38] A. R. Ali, ''Multi-dialect Arabic speech recognition,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–7.

[39] H. Mubarak, A. Hussein, S. A. Chowdhury, and A. Ali, ''QASR: QCRI Aljazeera speech resource—A large scale annotated Arabic speech corpus,'' 2021, *arXiv:2106.13000*.

[40] I. Hamed, P. Denisov, C.-Y. Li, M. Elmahdy, S. Abdennadher, and N. T. Vu, ''Investigations on speech recognition systems for low-resource dialectal Arabic–English code-switching speech,'' *Comput. Speech Lang.*, vol. 72, Mar. 2022, Art. no. 101278.

[41] H. Ahmed, H. Mamdouh, S. Ashraf, A. Ramadan, and M. Rashwan, ''RDI-CU system for the 2019 Arabic multi-genre broadcast challenge,'' *Education*, vol. 2019, 2019.

[42] F. S. Al-Anzi and D. AbuZeina, ''Synopsis on Arabic speech recognition,'' *Ain Shams Eng. J.*, vol. 13, no. 2, Mar. 2022, Art. no. 101534.

[43] A. Hussein, S. Watanabe, and A. Ali, ''Arabic speech recognition by end-to-end, modular systems and human,'' *Comput. Speech Lang.*, vol. 71, Jan. 2022, Art. no. 101272.

[44] A. Pervaiz, F. Hussain, H. Israr, M. A. Tahir, F. R. Raja, N. K. Baloch, F. Ishmanov, and Y. B. Zikria, ''Incorporating noise robustness in speech command recognition by noise augmentation of training data,'' *Sensors*, vol. 20, no. 8, p. 2326, Apr. 2020.

[45] N. Desai, K. Dhameliya, and V. Desai, ''Feature extraction and classification techniques for speech recognition: A review,'' *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 12, pp. 367–371, 2013.

[46] M. P. Kesarkar and P. Rao, ''Feature extraction for speech recognition,'' ESEDIB, 2003.

[47] S. B. Davis and P. Mermelstein, ''Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,'' *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[48] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[49] M. S. Masekwameng, T. B. Mokgonyane, T. I. Modipa, M. J. Manamela, and M. M. Mogale, ''Effects of language modelling for Sepedi-English code-switched speech in automatic speech recognition system,'' in *Proc. Int. Conf. Artif. Intell., Big Data, Comput. Data Commun. Syst. (icABCD)*, Aug. 2020, pp. 1–5.

[50] N. F. Hmad, *Deep Neural Network Acoustic Models for Multi-Dialect Arabic Speech Recognition*. Nottingham, U.K.: Nottingham Trent Univ., 2015.

[51] N. Tomashenko, ''Speaker adaptation of deep neural network acoustic models using Gaussian mixture model framework in automatic speech recognition systems,'' School Le Mans, 2017.

[52] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, ''The subspace Gaussian mixture model—A structured model for speech recognition,'' *Comput. Speech Lang.*, vol. 25, no. 2, pp. 404–439, 2011.

[53] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, ''Subspace Gaussian mixture models for speech recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4330–4333.

[54] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, ''Learning representations by back-propagating errors,'' *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[55] F. Richardson, D. Reynolds, and N. Dehak, ''Deep neural network approaches to speaker and language recognition,'' *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1671–1675, Oct. 2015.

[56] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ''Gradient-based learning applied to document recognition,'' *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[57] D. H. Hubel and T. N. Wiesel, ''Binocular interaction in striate cortex of kittens reared with artificial squint,'' *J. Neurophysiol.*, vol. 28, no. 6, pp. 1041–1059, Nov. 1965.

[58] H. Alsayadi, A. Abdelhamid, I. Hegazy, and Z. Taha, ''Data augmentation for Arabic speech recognition based on end-to-end deep learning,'' *Int. J. Intell. Comput. Inf. Sci.*, vol. 21, no. 2, pp. 50–64, Jul. 2021.

[59] V. Passricha and R. K. Aggarwal, ''A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition,'' *J. Intell. Syst.*, vol. 29, no. 1, pp. 1261–1274, Mar. 2019.

[60] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, ''Convolutional neural networks for speech recognition,'' in *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[61] M. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, ''Deep convolutional neural networks for LVCSR,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 8614, May 2013, pp. 8614–8618.

[62] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, ''Phoneme recognition using time-delay neural networks,'' *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 3, pp. 328–339, Mar. 1989.

[63] L. Bottou, F. F. Soulié, P. Blanchet, and J. S. Liénard, ''Speaker-independent isolated digit recognition: Multilayer perceptrons vs. dynamic time warping,'' *Neural Netw.*, vol. 3, no. 4, pp. 453–465, Jan. 1990.

[64] L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel, Y. LeCun, U. A. Müller, E. Sackinger, P. Simard, and V. Vapnik, ''Comparison of classifier methods: A case study in handwritten digit recognition,'' in *Proc. 12th IAPR Int. Conf. Pattern Recognit.*, vol. 2, 1994, pp. 77–82.

[65] X.-H. Le, H. V. Ho, G. Lee, and S. Jung, ''Application of long short-term memory (LSTM) neural network for flood forecasting,'' *Water*, vol. 11, no. 7, p. 1387, Jul. 2019.

[66] H. A. Alsayadi, A. A. Abdelhamid, I. Hegazy, and Z. T. Fayed, ''Arabic speech recognition using end-to-end deep learning,'' *IET Signal Process.*, vol. 15, no. 8, pp. 521–534, Oct. 2021.

[67] S. Zhou, L. Dong, S. Xu, and B. Xu, ''Syllable-based sequence-to-sequence speech recognition with the transformer in Mandarin Chinese,'' 2018, *arXiv:1804.10752*.

[68] S. Zhou, L. Dong, S. Xu, and B. Xu, ''A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on Mandarin Chinese,'' in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2018, pp. 210–220.

[69] L. Dong, S. Xu, and B. Xu, ''Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5884–5888.

[70] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, ''A comparative study on transformer vs RNN in speech applications,'' in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 449–456.

[71] H. Jiang, ''Discriminative training of HMMs for automatic speech recognition: A survey,'' *Comput. Speech Lang.*, vol. 24, no. 4, pp. 589–608, Oct. 2010.

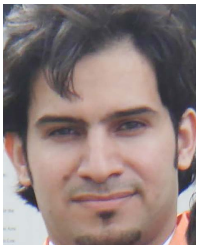
[72] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, ''FMPE: Discriminatively trained features for speech recognition,'' in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, Mar. 2005, p. I-961.

[73] Y. Hao, S. Mendelsohn, R. Sterneck, R. Martinez, and R. Frank, ''Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling,'' 2020, *arXiv:2009.03954*.

HAMZAH A. ALSAYADI received the B.Sc. degree from the Faculty of Sciences, Ibb University, Yemen, in 2007, and the M.Sc. degree from the Faculty of Computer and Information Technology, Cairo University, in 2016. His M.Sc. degree in Arabic named entity recognition. He is currently pursuing the Ph.D. degree with the Faculty of Computer and Information Sciences, Ain Shams University, Egypt. His research interests include speech recognition, speech synthesis, neutral language processing, data science, and deep learning.

ABDELAZIZ A. ABDELHAMID received the M.Sc. degree in computer science from the Faculty of Computer and Information Sciences, Ain Shams University, and the Ph.D. degree in computer engineering from the Faculty of Engineering, Auckland University, New Zealand. He is currently an Assistant Professor with the Department of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University. He is also working as an Assistant Professor with the Computer Science Department, College of Computing and Information Technology, Shaqra University. His research interests include speech and image processing and machine learning-based intelligent systems.

ISLAM HEGAZY received the B.Sc. and M.Sc. degrees from the Faculty of Computer and Information Sciences, Ain Shams University, and the Ph.D. degree from the University of Calgary, AB, Canada. He is currently an Assistant Professor with the Department of Computer Science, Faculty of Computer and Information Sciences. He has more than 18 years experience in research and teaching in many fields in the fields of computer science. His research interests include networks, security, and cloud computing. He acquired several managerial skills as the Director of the Scientific Computing Center, Ain Shams University, and the Coordinator of the Software Engineering Credit Hours Program, Faculty of Computer and Information Sciences.



BANDAR ALOTAIBI (Member, IEEE) received the Bachelor of Science degree (Hons.) in computer science (information security and assurance) from the University of Findlay, USA, the Master of Science degree in information security and assurance from Robert Morris University, USA, and the Ph.D. degree in computer science and engineering from the University of Bridgeport, USA. He is currently an Associate Professor with the Information Technology Department, University of Tabuk. His research interests include computer vision, network security, mobile communications, computer forensics, wireless sensor networks, and quantum computing.

ZAKI T. FAYED received the B.Sc. degree in electronic engineering (communication engineering) from the Military Technical College, in 1976, the M.Sc. degree in computer science (speech processing) from the Communication and Computer Department, Faculty of Engineering, Ain Shams University, in 1992, and the Ph.D. degree from the Communication Department, Faculty of Engineering, Ain Shams University, in 1997. He is currently a Professor with the Department of Computer Science, Faculty of Computer and Information Sciences. He has more than 30 years experience in research and teaching in many fields in the fields of computer science. His research interests include speech processing and computer security. He held several positions in Military IS Center, Ministry of Defense, Military Academy, Armed Forces, and Faculty of Computer and Information Sciences.

• • •