Arabic Speech Recognition using a Combined Deep Learning Model

Fathiyah Habeeb
Computer Science Department,
Libyan Academy, Tripoli, Libya
fathabibe@gmail.com

Abduelbaset Goweder
Computer Science Department
Libyan Academy, Tripoli, Libya
agoweder@academy.edu.ly

Abstract— ***Speech recognition is a valuable tool in various industries; however, achieving high accuracy remains a major challenge, despite the rapid growth of the speech recognition market. Arabic in particular lags behind other languages in the field of speech recognition, requiring further attention and development. To address this issue, this research uses deep neural networks to develop an automatic Arabic speech recognition model based on isolated words technology. A hybrid model, which is originally developed by Radfar et al. [1] for English speech recognition, is adopted and adapted to be used for Arabic speech recognition. This model combines the strengths of recurrent neural networks (RNNs), which are critical in speech recognition tasks, with convolutional neural networks (CNNs) to form a hybrid model known as ConvRNN. A specific model for Arabic speech recognition which is referred to as "Arabic_ConvRNN" model has been developed based on "ConvRNN" model. The adopted model is trained using an Arabic speech publicly available dataset of isolated words, along with a custom-generated dataset specially prepared for this research. The performance of the built model has been evaluated using standard metrics, including word error rate (WER), accuracy, precision, recall, and F-measure (also referred to as f1-score). In addition, K-fold cross-validation method has been employed to ensure robustness and generalizability. The results demonstrated that Arabic_ConvRNN model achieved a high accuracy rate of 95.7% on unseen data, with a minimal WER of about 4.3%. These findings highlight the model's effectiveness in accurately recognizing Arabic speech with minimal errors. Comparisons with similar models from previous studies further validated the superiority of Arabic_ConvRNN model. Overall, the Arabic_ConvRNN model shows great promise for applications requiring accurate and efficient Arabic speech recognition. This research contributes to narrowing the gap in Arabic speech recognition technology, offering a robust solution for accurately converting Arabic speech into text.***

Keywords—Arabic Speech Recognition, Neural Networks, Deep learning, convolutional neural networks (CNNs), Recurrent Neural Networks (RNNs).

I. INTRODUCTION

Nowadays, speech is still the most effective form of human communication, and speech recognition is one way to solve communication problems that arise with computers and other systems. However, getting it done has always been one of the hardest tasks. Over the past few decades, artificial intelligence (AI) has undergone rapid change, and one of the changes is the application of AI to speech processing. Using a computer program and its algorithms, automatic speech recognition converts a speech signal into a string of words. The primary goal of speech recognition is to allow machines to recognize and respond to sounds. It is the process by which a computer is able to "receive and interpret" speech and convert it into readable text or form. Automatic speech recognition (ASR) is the ability of a computer to understand speech as well as perform an action based on human instructions. In order to improve the performance of speech recognition systems, new algorithms have been developed after many years of research and development, some of which are based on neural networks.

One of the most important of these developments is the field of deep learning based on artificial neural networks that are similar to those in the human brain. Deep learning is "a set of methods that allow computers to learn from data without human supervision and intervention [2]. Today, deep learning is pervasive in our daily lives in the form of Google search engines, Apple's Siri, and Amazon and Netflix recommendation engines to name. Our ability to model human communication and interaction and enhance human-computer interaction is being expanded through deep learning techniques.

Some significant uses of this technology include recognizing commands, transcribing speech, translating foreign languages, and controlling security (such as verifying a person's identity for accessing services like telephone banking). ASR significantly simplifies and speeds up writing in computer applications compared to keyboard usage, and has the potential to assist individuals with disabilities in their social interactions. In addition, it has the capability to be utilized for the remote control of home lighting and electrical devices, allowing them to be turned on or off (Internet of Things).

DOI: 10.5281/zenodo.14553448

Speech recognition systems can be classified into different categories based on what type of utterances they can recognize. These categories are of four types: isolated words, connected words, connected speech, and spontaneous speech.
Automatic Speech Recognition systems consist of two stages:

1. A training phase in which the system learns the reference patterns that represent different speech sounds.
2. A recognition phase in which an unidentified speech signal is identified by comparing it to stored reference patterns.

The remaining of this paper presents a background, literature review, data collection, model building, model's results and performance, summary of the results, a comparison with other studies, conclusions, and future work.

II. Background

This section gives a brief background about Arabic language and speech recognition.

Arabic, a Semitic language, is among the most widely used and spoken languages in the world and the least known languages for speech recognition. An estimated number of speakers of Arabic is more than 313 million people with 270 million speakers as a second language. It is ranked fourth after Mandarin, Spanish and English [3]. Moreover, it is the language of the "Holy Quran" with 1.8 billion Muslims around the world in 2015 and is expected to increase to 3 billion in 2060 [4], with over 200 million native speakers across the Middle East and North Africa. In each Arabian community, there are two different types of the language that exist together. The choice of language used by an Arabic speaker mainly relies on sociolinguistic factors. Modern Standard Arabic (MSA) is the formal dialect. It is a well-organized form of the Arabic dialect spoken across the Arab world and is considered the "formal" version of the language. MSA is traditionally employed in literacy and education, and frequently utilized in spoken interactions where formality is stressed, such as TV documentaries, interviews, oral presentations, and ceremonial speeches [5]. Arabic has rich morphological and syntactic structures. It has complex grammar rules, and words can have different meanings based on context. Additionally, Arabic is written in a cursive script from right to left, and it has diacritical marks that change word meanings but are often omitted in written text. Arabic is spoken in many dialects that differ significantly across regions. Modern Standard Arabic (MSA) is the formal version used in writing and formal speech, but in everyday conversations, people speak in various local dialects, which can differ in pronunciation, vocabulary, and grammar.

Speech recognition is a technology that converts spoken language into written text. The primary goal is to enable machines, such as computers or voice-activated devices, to understand and interpret human speech. This technology has various applications, ranging from voice assistants and transcription services to hands-free control of devices. This field faces unique challenges due to the complexity of the Arabic language, its various dialects, and the differences in spoken and written forms. Fig. 1 shows a diagram of the components of a speech recognition system.

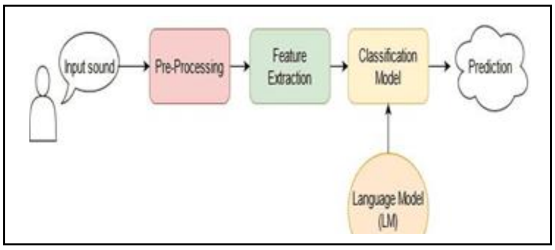


Fig. 1. A diagram of speech recognition system.

III. Literature Review

In this section, some Arabic speech recognition research is presented. Although Arabic is a widely spoken language, research conducted in the field of Arabic speech recognition is limited compared to other human languages. In this section, we take a look at previous work in the field of Arabic speech recognition. Most of work focuses on providing proper techniques for speech recognition and how to get good results using deep learning algorithms**.**

A. CNN and RNN Techniques

Rady et al. (2021) [6] developed a robust speaker-independent automatic Arabic speech recognition (AASR) model based on a convolutional neural network (CNN) for Arabic word recognition. The study leverages feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC) and Gammatone Frequency Cepstral Coefficients (GFCC), resulting in a significant improvement in recognition accuracy, achieving an impressive 99.77%. While the model's high performance is noteworthy, a more in-depth discussion of contributing factors, including data augmentation techniques, would enhance the analysis. Additionally, comparing the model's performance with other contemporary approaches, such as RNNs or transformers, could provide a more

DOI: 10.5281/zenodo.14553448

Alsayadi et al. (2021) [7] present an end-to-end approach for Arabic automatic speech recognition, employing novel modulation techniques that combine Connectionist Temporal Classification (CTC) with Convolutional Neural Network & Long Short Term Memory (CNN-LSTM) models. Their work highlights the superiority of CNN-LSTM with attention mechanisms over traditional models, achieving a 5.24% reduction in Word Error Rate (WER) compared to traditional methods and a 2.62% reduction compared to the CTC and attention combined model. However, the study uses a limited dataset of 7 hours of Standard Arabic speech, which may affect the generalizability of the results to other Arabic dialects and real-world settings. The paper also lacks detailed exploration of the system's performance in environments with noise and diverse dialects.

Alsayadi et al. (2022) [8] introduced an innovative CNN-LSTM model with an attention-based encoder-decoder framework for Dialectal Arabic Speech Recognition (DASR), a relatively unexplored area. Their model achieved a word error rate (WER) of 57.02% and a character error rate (CER) of 25.24%, demonstrating its ability to handle dialectal complexities. However, these results suggest that significant improvements are needed for the model to effectively recognize different dialects. The study's reliance on the Standard Arabic Single Speaker Corpus (SASSC) and Third Multi-Genre Broadcast (MGB-3) datasets may limit its generalizability across the diverse Arabic dialects. Additionally, the potential for overfitting due to limited data and the inherent challenges of dialectal variability highlight the need for further refinement.

Rafik et al. (2022) [9] presented an application of CNN combined with LSTM networks for Arabic speech recognition, achieving an accuracy of 88%. While this architecture effectively addresses the complexities of speech recognition, the accuracy is relatively modest, particularly given that the data was limited and tested in a noise-free, controlled environment. Moreover, the study could be strengthened by including more comparisons with current research in the field.

A novel approach was introduced by M. El Choubassi et al.(2018) [10] for developing an Arabic word recognition system using modular recurrent Elman neural networks (MRENN) which is a special kind of a Recurrent Neural Networks (RNN). This system offers an alternative to the traditional Hidden Markov Models (HMM)-based approach. The recognition accuracy of the system varies according to the speaker and the conditions, with one speaker achieving 100% accuracy in a clean environment, while another achieved 85% under different conditions. Although the overall accuracy differs across speakers and environments, the system is currently limited to recognizing a small vocabulary of only 6 words, which restricts the generalizability of the results and their applicability to wider contexts.

Abbas (2021) [11] tackled the significant challenge of developing a model capable of processing multiple Arabic dialects, which are highly diverse. The study implemented advanced deep neural networks, such as LSTMs and Gated Recurrent Units (GRUs), representing a modern approach to speech recognition. The model achieved a notable 14% error rate, surpassing earlier systems and highlighting its effectiveness. However, challenges remain, including limited resources for certain dialects, potential accuracy issues due to the removal of diacritics, and the complexity of training and deploying the deep network, which demands considerable computational power.

Abdelrahman et al.(2018) [12] presented Arabic speech recognition system using Recurrent Neural Networks (RNNs) without relying on dictionaries, making it a lexicon-free model. The model was tested on an extensive dataset from Al Jazeera, spanning ten years, which adds credibility to its performance. By utilizing the CTC algorithm, the model improves character recognition accuracy without the complexity of traditional Hidden Markov Models (HMMs), simplifying the training process. Although 1200 hours of Al Jazeera data were used, the dataset does not cover all Arabic dialects, which might affect the model's generalizability across different environments. While the system's independence from dictionaries is advantageous, it may struggle with recognizing uncommon or new words.

Naima et al.(2018) [13] present a comprehensive and advanced approach that utilizes Long Short-Term Memory (LSTM) networks for sequence learning, effectively addressing temporal sequence challenges. By employing a bidirectional LSTM model, the system is able to better recognize temporal patterns in audio signals, thereby enhancing overall performance. The use of Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction, a well-established and reliable technique in speech recognition, adds credibility to their results. However, the model is trained and evaluated on the "Spoken Arabic Digit Dataset," which is limited in scope and may not adequately capture the diversity of dialects and speech patterns in Arabic language, potentially limiting the generalizability of the findings. Additionally, the model's focus on recognizing only Arabic digits restricts its application, and the paper could benefit from discussing the potential for expanding the system to include a broader vocabulary or full sentences.

DOI: 10.5281/zenodo.14553448

B. *Other Deep Learning Techniques*

Yusra et al. (2016) [14] presented modern techniques such as Wavelet Transform and Fuzzy Neural Networks for Arabic speech recognition, enhancing the accuracy of isolated word recognition. It introduces innovative feature extraction methods by combining Linear Predictive Coding (LPC) with convolution, and integrating LPC with Wavelet Transform and Cepstral analysis, which contribute to the model's robustness. The study demonstrates high recognition accuracy, achieving 97.8% for trained words and 81.1% for untrained words. However, the dataset's limitation to 15 speakers and 10 words may affect the generalizability of the results. Additionally, the paper lacks a comprehensive comparison with previous research outcomes using alternative techniques.

Shareef and Irhayim (2021) [15] present a comprehensive survey of isolated Arabic word recognition, emphasizing various artificial intelligence techniques, which serves as a valuable resource for researchers new to this field. They effectively highlight the role of deep learning in addressing the complex patterns of Arabic speech. However, the paper could be improved by addressing dataset diversity limitations, expanding the scope to include more complex tasks, and providing a deeper analysis of the reviewed techniques.

Lallouani and Mohamed (2018) [16] presented the innovative use of Multitaper Frequency Cepstral Coefficients (MFCC-MT) and Gabor features for feature extraction, aiming to reduce spectral variance and enhance robustness against variability in speech signals. The study evaluates a comprehensive range of classification systems, including Continuous Hidden Markov Models (CHMM), Deep Neural Networks (DNN), and HMM-DNN hybrids, making it relevant for improving speech recognition performance in mobile networks, particularly through Distributed Speech Recognition (DSR) and Network Speech Recognition (NSR).

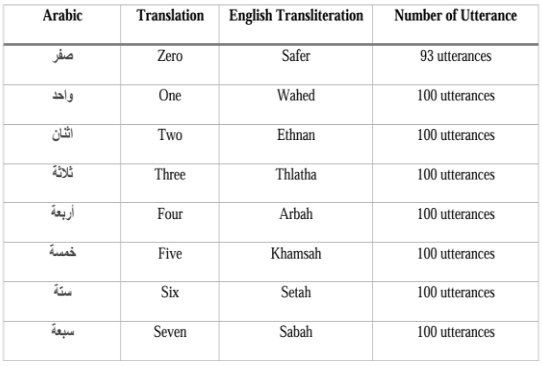
Wajdan (2019) [17] used Mel Frequency Cepstral Coefficients (MFCC) based on feature extraction and artificial neural network (ANN) classification method.

IV. Data Collection And Pre-Processing

This section describes the dataset that was used to train a hybrid model known as ConvRNN model.

An Arabic speech dataset for isolated words is adopted which is developed at the Department of Management Information Systems, King Faisal University. The dataset (referred to as Dataset1) comprises 9,992 instances of 20-word utterances spoken by 50 native male Arabic speakers. The recordings were captured at a sampling rate of 44,100 Hz and 16-bit resolution [18]. A sample of words of Dataset1 and their corresponding utterances are listed as shown in Table I.

TABLE I. A SAMPLE OF WORDS OF DATASET1.

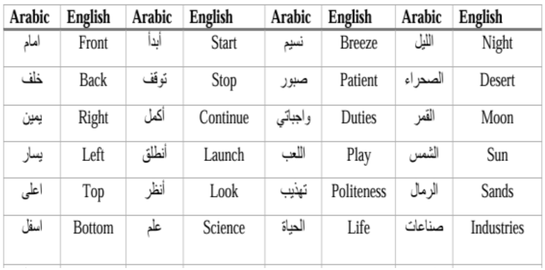


Arabic	Translation	English Transliteration	Number of Utterance
صفر	Zero	Safer	93 utterances
واحد	One	Wahed	100 utterances
اثنان	Two	Ethnan	100 utterances
ثلاثة	Three	Thlatha	100 utterances
أربعة	Four	Arbah	100 utterances
خمسة	Five	Khamsah	100 utterances
ستة	Six	Setah	100 utterances
سبعة	Seven	Sabah	100 utterances

Each file of Dataset1 is labeled using a specific coding system: S (Speaker Number).(Repetition Number).(Word Number). For example, "S01.01.01" represents the first speaker, the first recording, and the first word from the list of 20 words.

Additionally, new dataset (referred to as Dataset2) was recorded in WAV format, which included 80 new words spoken by 20 female Arabic speakers. A sample of words of Dataset2 and their corresponding utterances are listed as shown in Table II.

TABLE II. A SAMPLE OF WORDS OF DATASET2.



Arabic	English	Arabic	English	Arabic	English	Arabic	English
امام	Front	أبدأ	Start	نسيم	Breeze	الليل	Night
خلف	Back	توقف	Stop	صبور	Patient	الصحراء	Desert
يمين	Right	أكمل	Continue	واجباتي	Duties	القمر	Moon
يسار	Left	أنطلق	Launch	اللعب	Play	الشمس	Sun
اعلى	Top	أنظر	Look	تهذيب	Politeness	الرمال	Sands
اسفل	Bottom	علم	Science	الحياة	Life	صناعات	Industries

In addition, both datasets (Dataset 1 and Dataset 2) are combined in a single dataset called “Dataset 3” with a total size of 1.1 GB. By merging these two datasets, a more diverse set of training examples is obtained which enhances the ability of any built model to generalize and recognize a wider variety of speech patterns. The integration of both datasets provides a richer representation of the Arabic language, covering more variations in pronunciation, accents, and contextual usage. This larger dataset contributes to improving the model’s robustness and performance across different speech recognition tasks.

Pre-processing of speech data is a crucial stage that converts raw audio recordings into a format more

DOI: 10.5281/zenodo.14553448

suitable for machine learning algorithms. Pre-reprocessing audio data includes decoding, normalization, and feature extraction to make the data more consistent and easier to build voice recognition model.

Audio decoding is the process of converting audio data from its raw form into a suitable format (numeric data), such as floating point numbers that can be effectively processed and analyzed by machine learning models.

Normalization typically involves scaling the amplitude values of the waveform so that they fall within a specific range, such as [-1, 1] or [0, 1]. This is important because the amplitude values of audio waveforms can vary widely depending on factors such as recording equipment, microphone sensitivity, and environmental conditions. Normalizing the audio data ensures that any variations in volume or amplitude do not negatively impact the performance of the Speech Recognition models.

Common techniques for extracting features from audio data include: MFCCs, spectrograms, LPC, pitch & formant frequencies, and power and zero-crossing ratio. Spectrogram technology was used for feature extraction in this study, which involves converting audio signals into spectrograms, which are visual representations of the spectrum of signal frequencies that change over time A spectrogram is generated using a sliding time window in which a short-time Fourier transform (STFT) is performed. As a time-frequency visualization of a speech signal. Fig. 2 shows Voice signal for the word "Atansheet التنشيط" and feature extraction using spectrogram.

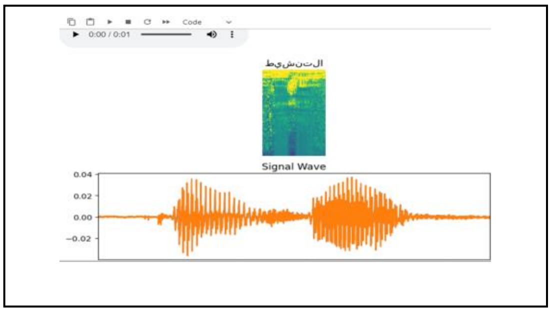


Fig. 2. Voice signal for the word "التنشيط" and feature extraction using spectrogram.

In this section, the main components of the adopted model used for Arabic speech recognition are discussed. In this model, CNN and RNN algorithms are combined to produce the ConvRNN model which will be used to recognize isolated words in Arabic language. This hybrid model leverages the strengths of both CNNs and RNNs. CNNs are excellent at capturing spatial hierarchies in data. For audio signals, CNNs can effectively identify patterns in the spectrograms that represent different phonetic elements. RNNs are designed to handle sequential data, making them ideal for capturing temporal dependencies in audio signals. The ConvRNN model thus combines the feature extraction capabilities of CNNs with the sequence modeling capabilities of RNNs to achieve high accuracy in recognizing isolated Arabic words. The model which is planned to be built is shown in Fig. 3 which illustrates a diagram of ConvRNN model. It represents a schematic diagram of a deep learning model architecture, specifically designed for tasks like speech recognition. The description of each component of the diagram is as follows: the model begins with a Convolutional Neural Network (CNN) layer to extract local features such as edges and textures. A second CNN layer follows, refining these features to capture more complex patterns. The data is then reshaped to fit the subsequent layers, specifically a Bidirectional GRU (Gated Recurrent Unit) layer, which captures temporal dependencies by processing the sequence in both forward and backward directions which is crucial for tasks like speech recognition. A Dropout Layer is applied to prevent overfitting, followed by a Dense Layer that combines the learned features for final classification. The model concludes with a Softmax layer that converts outputs into probabilities, identifying the most likely class. Each component contributes to the model's ability to effectively process and recognize patterns, leading to accurate predictions. Fig. 4 shows a diagram of CNN part of the ConvRNN model. In this diagram, two layers of (2D CNN) will be used. Each layer will be followed by batch normalization and ReLu activation.

The RNN part of the ConvRNN model is responsible for handling the temporal dependencies in the input data after the convolutional layers have

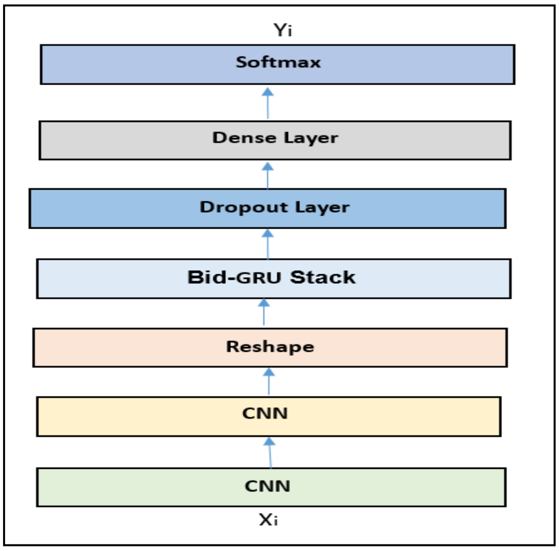


Fig. 3. A diagram of the ConvRNN model.

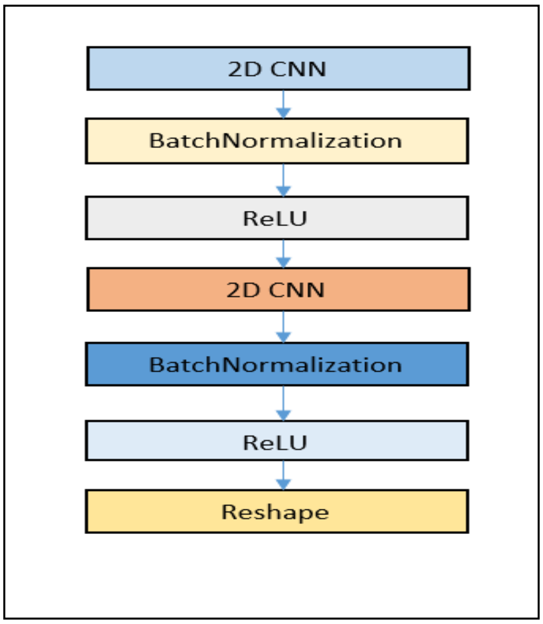


Fig. 4. A diagram of CNN model.

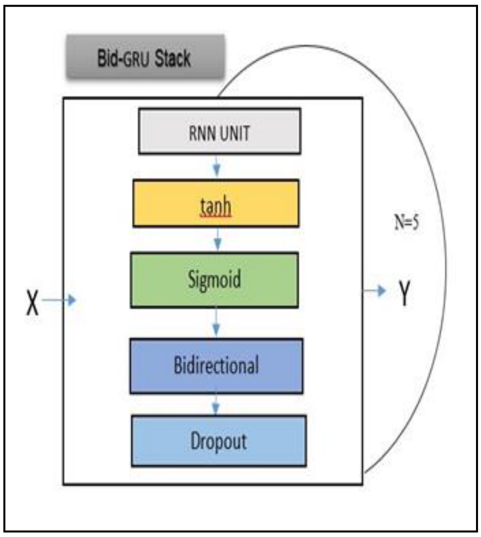


Fig. 5. A diagram of RNN model.

extracted spatial features. Fig. 5 illustrates a diagram of RNN model and how the layers are structured. The diagram represents a Bidirectional Gated Recurrent Unit (Bid-GRU) Stack, which includes several layers working together to process input data, denoted as X, and produce an output, Y. At the core of the structure is the RNN Unit, specifically a Gated Recurrent Unit (GRU), responsible for handling sequential data where each input depends on the preceding one. After the RNN unit processes the data, a tanh activation function is applied, which introduces non-linearity by mapping values to a range between -1 and 1. This is followed by a Sigmoid activation, which further refines the data by mapping it to a range between 0 and 1, a step commonly used for binary decisions or scaling inputs. The next component is the Bidirectional layer, which processes the input sequence in both forward and backward directions, allowing the model to capture dependencies from both the past and future within the sequence. This makes the model more effective in understanding complex patterns. To ensure robustness and prevent overfitting, a Dropout layer is added, which randomly deactivates a portion of the neurons during training to force the model to learn more generalized patterns. The process is repeated for N=5 iterations, refining the data through multiple passes before producing the final output.

VI. MODEL'S RESULTS AND PERFORMANCE

This section presents the obtained results of training Arabic_ConvRNN model using three different datasets namely (Dataset1, Dataset2, and Dataset3) using two methods: without cross-validation and with cross-validation. The key metrics considered are accuracy, and word error rate (WER) across three different datasets.

A. Without Cross-Validation Method

Evaluating the performance of Arabic_ConvRNN model using three different datasets has revealed that the built models have achieved impressive accuracies (99.7%, 90%, and 99%) for (Dataset1, Dataset2, and Dataset3) respectively. Also, the average Word Error Rate (WER) is about (0.5%, 10%, and 0.88%) for (Dataset1, Dataset2, and Dataset3) respectively. These results indicate that the built models accurately recognized words with minimal errors. A lower WER reflects better performance in recognizing and transcribing spoken words. These results have showed that the Arabic_ConvRNN model is highly effective in recognizing Arabic speech.

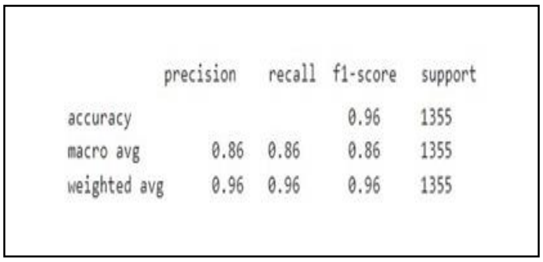
B. With Cross-Validation Method

Using 10-fold cross-validation method with Arabic_ConvRNN model resulted in a word error rate (WER) of 4.3%. That is, 4.3% of the words predicted by the model differ from the actual words in the reference texts. This low error rate indicates

DOI: 10.5281/zenodo.14553448

that the model is skilled at accurately transcribing spoken words into text across different folds of the data set. Achieving this low level of WER is critical for applications that require accurate speech recognition, such as voice assistants or transcription services, where accuracy directly impacts ease of use and user satisfaction. With an accuracy of 95.7%, the model correctly identifies 95.7% of the words in the predicted texts compared to the reference texts. This high accuracy emphasizes the model's ability to effectively capture and interpret spoken language, and translate it into accurate textual representations. High accuracy is essential in scenarios where correct interpretation of spoken words is crucial, ensuring reliable communication and interaction in different applications.

Arabic_ConvRNN model (using 10-fold cross-validation method) is overall evaluated on the Dataset3 using the confusion matrix. The average of the overall metrics (precision, recall, and f1-score) is about 0.86, which reflects the average performance of the model in all categories. The weighted average is higher, at 0.96, due to the higher number of cases correctly classified in larger categories. Fig. 6 shows the results of the classification report for Dataset3 using 10-fold cross-validation method.



```
              precision    recall  f1-score   support

    accuracy                           0.96      1355
   macro avg       0.86      0.86      0.86      1355
weighted avg       0.96      0.96      0.96      1355
```

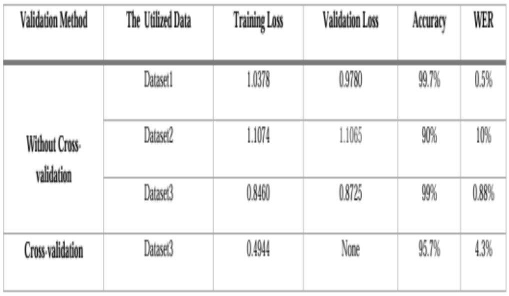
Fig. 6. The classification report for Dataset3 using 10-fold cross-validation method.

Overall, the model performs well on all datasets, with the highest accuracy achieved on Dataset 1 without cross-validation. However, the use of cross-validation on Dataset 3 helped reduce overfitting and resulted in a more reliable WER of 4.3%, indicating that cross-validation plays a crucial role in enhancing the model's generalizability and robustness.

VII. SUMMARY OF THE RESULTS

The obtained results of training Arabic_ConvRNN model using three different datasets namely (Dataset1, Dataset2, and Dataset3) are summarized as given in Table III. Without cross-validation method, it is obvious that Arabic_ConvRNN model using three different datasets has achieved impressive accuracies (99.7%, 90%, and 99%) for (Dataset1, Dataset2, and Dataset3) respectively. Also, the average Word Error Rate (WER) of (0.5%, 10%, and 0.88%) for (Dataset1, Dataset2, and Dataset3) respectively indicates that the model accurately recognized words with minimal errors. A lower WER reflects better performance in recognizing and transcribing spoken words. These results have showed that Arabic_ConvRNN model is highly effective in recognizing Arabic speech. On the other hand (using cross-validation method), it can be observed that Arabic_ConvRNN model using the entire data (Dataset3) has achieved an accuracy of about 95.7% which is considered excellent figure. Also, the average Word Error Rate (WER) of about 4.3% indicates that the model is highly performing in recognizing Arabic spoken words. Based on these results, it can be concluded that Arabic_ConvRNN model consistently performs well, with low error rates in recognizing Arabic spoken words indicating good generalization ability.

TABLE III. SUMMARY OF THE OBTAINED RESULTS FOR ARABIC_CONVRNN MODEL USING THREE DIFFERENT DATASETS.



Validation Method	The Utilized Data	Training Loss	Validation Loss	Accuracy	WER
Without Cross-validation	Dataset1	1.0378	0.9780	99.7%	0.5%
	Dataset2	1.1074	1.1065	90%	10%
	Dataset3	0.8460	0.8725	99%	0.88%
Cross-validation	Dataset3	0.4944	None	95.7%	4.3%

VIII. A COMPARISON

In this section, the obtained results of this research are compared with those obtained by previous studies. To compare the performance of Arabic_ConvRNN model with other similar models in the field of Arabic speech recognition, the results of different previous studies that used different deep learning techniques will be explored. The performance metrics and datasets used in these studies provide a comprehensive understanding of how Arabic_ConvRNN model stands up relative to existing models.

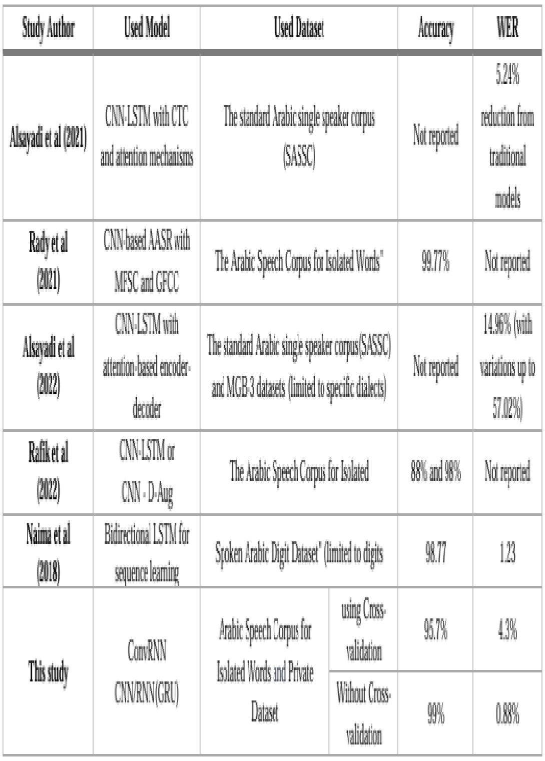
Arabic_ConvRNN model has demonstrated outstanding performance in the recognition of isolated Arabic speech, achieving up to 99% accuracy on unseen data and a word error rate (WER) as low as 0.8%. These results underscore the model's strong capability in accurately recognizing obtained without the use of cross-validation, relying instead on a conventional data split. When cross-

DOI: 10.5281/zenodo.14553448

validation was applied, the model maintained high performance, achieving a 95.7% accuracy and a WER of 4.3%.

In comparison to other models, Arabic_ConvRNN model shows encouraging and promising results. It outperforms the CNN-LSTM/CTC model [15] and the CNN-LSTM with attention-based model [16] in terms of WER. However, it is important to note that these previous models were trained on different dialects of the Arabic language, while Arabic_ConvRNN model was trained exclusively on Modern Standard Arabic (MSA). When compared to the CNN/MFCC/GFCC model [14], which also achieved high accuracy using a similar data approach, Arabic_ConvRNN model's accuracy is definitely comparable. Additionally, Arabic_ConvRNN model exhibits superior performance in isolated Arabic speech recognition compared to the CNN combined with LSTM model [17] when evaluated using the same data. Table IV compares the obtained results of this study with those of some previous studies.

TABLE IV. A COMPARISON OF THIS STUDY WITH OTHER STUDIES.



Study Author	Used Model	Used Dataset		Accuracy	WER
Alsayadi et al (2021)	CNN-LSTM with CTC and attention mechanisms	The standard Arabic single speaker corpus (SASSC)		Not reported	5.24% reduction from traditional models
Rady et al (2021)	CNN-based AASR with MFSC and GFCC	The Arabic Speech Corpus for Isolated Words"		99.77%	Not reported
Alsayadi et al (2022)	CNN-LSTM with attention-based encoder-decoder	The standard Arabic single speaker corpus(SASSC) and MGB-3 datasets (limited to specific dialects)		Not reported	14.96% (with variations up to 57.02%)
Rafik et al (2022)	CNN-LSTM or CNN - D-Aug	The Arabic Speech Corpus for Isolated		88% and 98%	Not reported
Naima et al (2018)	Bidirectional LSTM for sequence learning	Spoken Arabic Digit Dataset" (limited to digits		98.77	1.23
This study	ConvRNN CNN/RNN(GRU)	Arabic Speech Corpus for Isolated Words and Private Dataset	using Cross-validation	95.7%	4.3%
			Without Cross-validation	99%	0.88%

IX. Conclusions

Natural language processing applications often face numerous hurdles, primarily because of the complexity of language and the challenge of dealing with textual data rather than numerical data. Many algorithms and libraries have been created to assist in training deep learning models on text data and discovering patterns in text processing applications. In this study, we explored the application of ConvRNN model for Arabic speech recognition, a field that has seen significant advancements but still faces unique challenges due to the complexity and diversity of Arabic language. ConvRNN model combines convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to take their advantages in capturing spatial and temporal patterns in speech data. The evaluation of Arabic_ConvRNN model has been conducted using a comprehensive approach, incorporating both publicly available datasets and a custom-generated dataset. The model's performance has been assessed using standard metrics such as precision, recall, f1-score, and Word Error Rate (WER). In addition, employing K-fold cross-validation method to ensure robustness and generalizability. The results demonstrated that Arabic_ConvRNN model achieved a high accuracy rate of 95.7% on unseen data, with a minimal WER of just 4.3%. These findings highlight the model's effectiveness in accurately recognizing Arabic speech with minimal errors.

X. FUTURE WORK

The promising obtained results using ConvRNN model highlight its potential for distinguishing Arabic speech across various scenarios, pointing to its future development and application. The creation of a graphical user interface (GUI) to facilitate user interaction with the model is a crucial step towards broadening its accessibility.

Expanding and diversifying the training datasets is expected to enhance the model's robustness and generalization capabilities. Incorporating a wider range of dialects, accents, and contextual variations will likely improve the model’s performance in real-world applications.

Furthermore, adapting ConvRNN model for real-time applications, such as voice assistants or automated transcription services, presents an opportunity for practical deployment. It is essential to ensure that the model can handle various background noise conditions and adverse environments to enhance its effectiveness in real-world applications.

ACKNOWLEDGMENTS

We first and foremost praise and give thanks to Allah for making this research possible. In addition, we express our appreciation to those who willingly contributed to the creation of a specifically prepared dataset (Dataset2) for this research.

For the logistical support, we would like to thank Libyan Academy and Libyan Ministry of Higher Education & Scientific Research.

DOI: 10.5281/zenodo.14553448

REFERENCES

[1] Radfar. M . Barnwal. R . Swaminathan. R. Chang. F. Strimel. G. Susanj. N & Mouchtaris. A. (2022*). "ConvRNN-T: Convolutional augmented recurrent neural network transducers for streaming speech recognition".*Alexa Machine Learning, Amazon, USA. doi.org/10.48550/arXiv.2209.14868

[2] Kamath.U. Liu. J & Whitaker. J. (2019). "*Deep learning for NLP and speech recognition". *McLean VA. USA. Nashville, TN. USA. doi.org/10.1007/978-3-030-14596-5_1

[3] El Choubassi. M. M. El Khoury. H. E. Alagha. C. E. J. Skaf, J. A & Al-Alaoui. M. A. (2004). *"Arabic speech recognition using recurrent neural networks*". In Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (pp. 543–547). IEEE. doi.org/10.1109/ISSPIT.2003.1341178

[4] Lipka. M & Hackett. C. (2017, April 6). *"Why Muslims are the world's fastest-growing religious group".* Pew Research Center.https://www.pewresearch.org/fact-tank/2017/04/06/why-muslims-are-the-worlds-fastest-growing-religious-group/

[5] Al Waseem. M. (2021). "*Speech audiometry*: *Arabic word recognition test for adults*". (Doctoral dissertation, Kent State University College of Education. Health and Human Services). doi.org/10.21608/ejle.2020.47685.1015

[6] Rady. E. Hassen. A. Hassan. N & Hesham. M. (2021). *"Convolutional neural network for Arabic speech recognition*". Egyptian Journal of Language Engineering. 8(1). 1–15. doi.org/10.21608/ejle.2020.47685.1015

[7] Alsayadi. H. A. Abdelhamid. A. L. Hegazy. I & Fayed. Z. T. (2021). *"Arabic speech recognition using end-to-end deep learning".* IET Signal Processing. doi.org/10.1049/sil2.12057

[8] Alsayadi. A. H. Hagree. S. Abdelhamid. A. L & Alqasemi. F. (2022). *"Dialectal Arabic speech recognition using CNN-LSTM based on end-to-end deep learning*". Proceedings of the International Conference on Smart Technologies and Applications (pp. 42–56). doi.org/10.1109/eSmarTA56775.2022.9935427

[9] Rafik. A. Zouhaira. N. Salah. Z. Dhaou. B. Henri. N & Mounir. Z. (2022). "*Deep convolutional neural network for Arabic speech recognition*". In Proceedings of the International Conference on Natural Language Processing (pp. 120–134). doi.org/10.1007/978-3-031-16014-1_11

[10] El Choubassi. M. El Khoury. H. Jabra. C. Skaf. J. & Al-Alaoui M. (2018). *"Arabic speech recognition using recurrent neural networks*". In Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology. https://doi.org/10.1109/ISSPIT.2003.1341178

[11] Abbas. R. (2021). *"Multi-dialect Arabic speech recognition*". In Proceedings of the IEEE International Joint Conference on Neural Networks (pp. 605–616). doi.org/10.1109/IJCNN48.605.2020.9206658

[12] Abdelrahman.A. Yasser. H. Khaled. S & Sergio. T. (2018). *"End-to-end lexicon-free Arabic speech recognition using recurrent neural networks".* In Proceedings of the International Conference on Artificial Intelligence and Applications (pp. 315–324). doi.org/10.1142/9789813229396_0011

[13] Naima. Z. Samir. A. Hassen. B & Christian, R. (2018). *"Bi-directional recurrent end-to-end neural network classifier for spoken Arab digit recognition*". Proceedings of the 2nd International Conference on Natural Language and Speech Processing. doi.org/10.1109/ICNLSP.2018.8374374

[14] Yusra. F & Maher. K. (2016). *"Speech recognition of isolated Arabic words via using wavelet transformation and fuzzy neural network*". Computer Engineering and Intelligent Systems. 7(3). 1–10. ISSN 2222-1719 (Paper). ISSN 2222-2863.

[15] Shareef. R & Irhayim.Y. F. (2021). *"A review: Isolated Arabic words recognition using artificial intelligent techniques".* Journal of Physics: Conference Series, 1897, 012026. doi.org/10.1088/1742-6596/1897/1/012026

[16] Lallouani. B & Mohamed. D. (2018). "*Improving continuous Arabic speech recognition over mobile networks using DSR and NSR with MFCCs features transformed*". International Journal of Circuits. Systems and Signal Processing, 12(2), 42–56.

[17] Wajdan. A. Sarah. A. Noura. A & Anfal. A. (2019). *"Arabic speech recognition with deep learning: A review*". In Proceedings of the International Conference on Arabic Language Processing (pp. 15–31). doi.org/10.1007/978-3-030-21902-4_2

[18] Department of Management Information Systems. King Faisal University. (2014). *"The Arabic speech corpus for isolated words".* Retrieved from https://www.cs.stir.ac.uk/~lss/arabic/