ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Kurdish Speech to Text Recognition System Based on Deep Convolutional-recurrent Neural Networks

Lana Sardar Hussein, Sozan Abdulla Mahmood

Department of Computer Science, College of Science, University Sulaimanyah, Sulaimanyah, Kurdistan Region, Iraq

ABSTRACT

In recent years, deep learning has had enormous success in speech recognition and natural language processing. In other languages, recent progress in speech recognition has been quite promising, but the Kurdish language has not seen comparable development. There are extremely few research papers on Kurdish speech recognition. In this paper, investigated Gated Recurrent Units (GRUs) which is one of the popular RNN models to recognize individual Kurdish words, and propose a very simplified deep-learning architecture to get more efficient and high accuracy model. The proposed model consists of a combination of CNN and GRU layers. The Kurdish Sorani Speech KSS dataset was created for the speech recognition system, as its 18799 sound files for 500 formal Kurdish words. Finally, the model proposed was trained with collected data and yielded over %96 accuracy. The combination of CNN an RNN (GURs) for speech recognition achieved superior performance compared to the other feed-forward deep neural network models and other statistical methods.

Index Terms: Deep Learning, Gated Recurrent Units, Kurdish Speech Recognition, Convolutional Neural Network

1. INTRODUCTION

Speech is a natural means for people to communicate with one another. Automatic Speech Recognition (ASR) is the technique through which a computer can recognize spoken words and understand what they are saying. The ASR is the first component of a smart system. It is a method of converting an auditory signal into a string of words, which can then be used as final outputs or inputs in natural language processing. The purpose of ASR systems is to recognize human-spoken natural languages. ASR technology is commonly utilized in computers with speech interfaces, foreign language applications, dictation, hands-free operations and controls, and other features that enable interactions between machines and humans faster and easier than using keyboards [1].

ASRs are designed using a variety of methodologies, the most notable of which is the Hidden Markov Model (HMM) and machine learning-based methods, such as Artificial Neural Networks (ANNs) and Convolutional Neural Network (CNN) [2]. Increasing the accuracy and efficiency of these systems is one of the issues that still exist in this sector. Deep learning, a relatively new technology, has been widely employed to address this issue. Because an audio signal is a sample of sequential data, meaning its present value is reliant on all past values, in this work RNN (GRU) applied in addition with CNN. RNN is a type of artificial neural network. It involves a sequential data connection with the hidden neurons. It can be applied for the applications of text, audio, and video. It deals with sequential data from the

Access this article online

DOI:10.21928/uhdjst.v6n2y2022.pp117-125

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Lana Sardar Hussein, Department of Computer Science, College of Science, University of Sulaimanyah, Sulaimanyah, Kurdistan Region, Iraq. lana.salih@univsul.edu.iq

Received: 29-04-2022 Accepted: 07-09-2022 Published: 18-11-2022

analyzed the sequence at each time depending on the previous time in a directed cycle. LSTM units and Gated Recurrent Units (GRUs) are variations type of RNN. Thus, Recurrent Neural Networks (RNNs) are employed for processing speech signals [3].

Kurdish is an Indo-Iranian branch of Indo-European languages that are spoken by about 40 million People in Western Asia, primarily in Iraq, Turkey, Iran, Syria, Armenia, and Azerbaijan [3]. Kurdish contains several dialects, as well as its grammatical system and extensive vocabulary [4], [5].

Central Kurdish (also known as Sorani) and Northern Kurdish are the two most widely spoken dialects of Kurdish (also called Kurmanji). Zazaki and Gorani are two further dialects spoken by smaller groups (also known as Hawrami). Kurmanji is the Kurdish language spoken in northern Kurdistan (in Turkey, Syria, and northern Iraq) and written in the Latin (Roman) alphabet; it is also supported by Google Speech Recognition. The Sorani dialect is spoken primarily in the southeast, including Iran and Iraq, and is written in a modified variant of the Arabic alphabet. There is no data for the Sorani dialect in Google Speech Recognition [6].

In [7] mention that, Kurdish is hampered by a lack of resources to support its computational processing needs. Only a few attempts to develop voice recognition resources for the Kurdish language have been made thus far, necessitating the creation of a dataset for their research.

The major contribution of this work is design and implementation of a straightforward hybrid speech to text model for Kurdish (Sorani) that comprises three CNN layers and three (GRUs) layers, this combination in the proposed model architecture produced results that were more accurate.

The rest of this paper is organized as follows: Section two reviews the related works. Section three is the data collections workflow. Section four presents the model architecture and proposed method. In Section five, results are discussed, and finally, the conclusion is in Section 6.

2. LITERATURE REVIEW

Few attempts have been made to recognize Kurdish speech, this review focused on first: those papers in low resources languages (Arabic, Persian and Kurdish), and how audio datasets are built/collected with. Second: CNN and RNN techniques used for recognition is concerned. Kurdish character recognition has received some recent research such as [8]-[10]; however, our work on speech recognition is still in its early stages. The first attempts for Kurdish speech recognition in [7] which presents a dataset extracted from Sorani Kurdish texts from grades one to three of primary school in Iraq's Kurdistan Region. The first attempts for Kurdish speech recognition in [7] which presents a dataset extracted (BD-4SK-ASR) from Sorani Kurdish texts from grades one to three of primary school in Iraq's Kurdistan Region, which contains 200 sentences. Using CMUSphinx to create ASR, narrated by a single speaker using Audacity software at a sampling rate of 16000 and a 16-bit rate mono single channel. After that, another attempts for Kurdish language arise in [11] created a dataset for their work in Kurdistan, Iran, and used Kalditoolkit to develop the identification engine with SGMM and DNN algorithms for the aquatic model. The authors presented WER of Jira ASR system for different topics (SGMM model trained and evaluated by Office data) which are (General: 4.5%, Sport: 10.0%, Economic: 10.9, Conversation: 11.6%, Letter: 11.7%, Politics: 13.8%, Social: 15.3%, Novel: 16.0%, Religious: 16.2%, Scientific/Technology: 17.1%, and Poet: 25.2).

For isolated word recognition, some Arabic papers been reviewed. In [12] proposed, an Arabic digit classification system using 450 Arabic spoken digits. Based on a speaker-independent system, the accuracy was around 93%, the system is based on combining wavelet transform with linear prediction coding LPC method to extract the feature and the probabilistic neural network PNN for classification.

The work by [13] employed Sphinx technologies to recognize solitary Arabic digits with data provided by six different speakers. The system achieved an 86.66%-digit recognition accuracy, examine the use of a distributed word representation and a neural network for Arabic speech recognition. Furthermore, the neural network model allows for robust generalization and improves the ability to combat data sparseness. The inquiry approach also comprises a variety of neural probabilistic model configurations, an n-gram order parameter experiment, output vocabulary, normalization method, model size, and parameters. The experiment was carried out on Arabic news and discussion broadcasts.

Then, in [14] utilized an LSTM neural network for frame-wise phoneme classification on the Farsdat data set, and in [15], they employed a DLSTM with a CTC output layer for Persian phoneme recognition on the same data set.

The rest of this review focused on papers that used CNN, RNN, and GRU or combining of these techniques.

In [16] a significant study was reported. The authors utilized a deep Recurrent Neural Networks (RNN) model that was end-to-end with appropriate regularization. On the TIMIT phoneme recognition benchmark, they found that RNN, namely, Long Short-Term Memory (LSTM), had a test error of 17.7%.

There are, nevertheless, several studies underway to build computational tools for the Kurdish language. In [15] collects a tiny corpus named corpus of contemporary Kurdish newspaper texts (CCKNT), which contains 214K Northern Kurdish dialect terms. Pewan text corpus for Central Kurdish and Northern Kurdish was collected from two online news organizations. The Pewan corpus contains around 18 million tokens for the Central Kurdish dialect and approximately 4 million tokens for the Northern Kurdish dialect. This corpus serves as a validation set for information retrieval applications [17].

In [18], they offered a speech-to-text conversion strategy for the Malayalam language that employs deep learning techniques. For the training, the system is looking at 5–10 solitary words. Mel-frequency cepstral coefficients are acquired for the preprocessing phase. HMM is used to identify the speech and training after the preprocessing, syllabification, and feature extraction procedures. The LSTM was used to construct a speech recognition system based on ANN. The system has a 91% accuracy.

A recurrent neural network approach called LSTM to distinguish individual Bengali words was used in [19] The model is a two-layer deep recurrent neural network with 100 LSTM cells in each layer, 30 unique phonemes are detected, the last layer is a SoftMax output layer with 30 units, and the data set was used with a total of 2000 words. Fifteen different male speakers contributed to the audio speeches. Making a 75:12.5:12.5 split of the dataset for training, validation, and testing purposes. The test run yielded a phoneme detection error rate of 28.7% and a word detection error rate of 13.2%.

In [20] revised standard GRUs for phoneme recognition Purposes and proposed that Li-GRU architecture is a simplified version of a standard GRU, in which the reset gate is removed and ReLU activations are considered, this research worked with (TIMIT, DIRHA, CHiME, TED) corpus Li-GRU outperforms GRU in all the considered noisy environments, with achieving higher performance the bus (BUS) environment (the noisiest) relative improvement of 16% against the relative improvement of 9.5% observed in the street (STR). WER % calculated for DIRHA corpus in real part for (MFCC = 27.8, FBANK = 27.6, fMLLR = 22.8).

The study in [21] employed LSTM and two datasets in this project for Arabic speech recognition. The 1-digit dataset consists of 8800 tokens with a sampling rate of 11025 Hz, and it was created by asking 88 Arabic native speakers to repeat all digits 10 times. 2-TV command dataset: 10000 tokens for 10 TV commands at a sampling rate of 16000 Hz are contained in this dataset; finally, the author reached over 96% accuracy.

The author proposed four different model structures for speech Emotion recognition in [22], which are Model-A (1D CNNs-FCNs), Model-B (1D CNNs-LSTM-FCNs), Model-C (1D CNNs-GRU-FCNs), and Ensemble Model-D, which combines Model-A, Model-B, and Model-C, adding LSTM, and GRU after CNN blocks in models B and C results in increased accuracy (TESS) In total, there are 2800 audio files with 200 target words. 2-RAVDESS audio files have a resolution of 1440 pixels and a sample rate of 48 kHz. 3-SAVEE has a total of 1920 samples. 4-EMO-DB Berlin it contains 535 German-language audio recordings. CREMA-D is the fifth step in the CREMA-D process. It makes use of 7442 records.

3. DATA COLLECTION

This section will discuss how to collect data and which type of data should be collected. Those data were gathered through official administration papers in the University of Sulaimani College of Science, which totaled 500 different words and were collected by 30 speakers, 13 are female and 17 are male.

3.1. Kurdish Sorani Speech (KSS) Dataset

As mentioned before, there is no available dataset in Kurdish Sorani, which lead us to make our dataset for this research work here are the details of workflows, choosing individual words from the governmental worksheet, and arranging them in 30 to 50 words in each paragraph. Four hundred words were read by 30 volunteers and the last 100 words were read by two different male and female readers; the total number of words reached 500 words. There were 30 volunteers in total, with 14 males and 16 females, 9 from family, and 21 from universities. The volunteers' ages ranged from 20 to 40.

The volunteer was asked to read the paragraph as individual words, which means between each word makes silence for at least 0.2 s. Some speakers were asked to read each paragraph 1 time, which leads to 2–3 min' duration of each file, but some others were asked to read each same word 3–5 times the duration of these files reached 5–7 min.

3.2. Recording Circumstance

KSS data sets were collected in two environments office and home with two recording devices that Table 1 – indicates all information needed for data collection.

The dataset consists of 500 words of numeric numbers and formal words. However, from "١١" in Sorani reading "یازده" or "یانزه" in Latin reading "Yazde" or "Yanze" which is eleven in numeric number, to "٢٠" in Sorani reading "بیست" in Latin reading "Bîst" which is twenty numeric number, and also for (١٠٠٠،١٠٠،٩٠،٨٠،٧٠،٦٠،٥٠،٤٠،٣٠ ،١٠٠٠٠٠٠) as the same above for each numeric number, in total 41 tokens, the rest of 461 words containing formal words, weekdays, Kurdish month name, Kurdish pronoun, prefixes, and suffixes, and there are some words in the Kurdish language used to join two same or different words or sentences like (وە ، و ، کە، ی، یان، بۆ، لە), Table 2 shows part of the dataset.

3.3. Recording Technology

This section will discuss the property of the application that is used for recording speech and the recording conditions being explained.

For recording sounds in an office environment, Audacity was utilized since it is a free, easy-to-use, multi-track audio editor and recorder for Windows, MAC OS, GNU/Linux, and other operating systems that can also export sound files as MP3s (MP3, OGG, and WAV).

Using this application, need to set up some recording conditions, that could work properly with the deep learning model, these conditions are described below. using both mono recording channel and stereo recording channel and sample rate of 16000 Hz as it is near to the normal human sound and also using 44100 Hz after converting it to 16000 Hz.

TABLE 1: The information on data collection

Title	Value
Dataset name	Kurdish Sorani Speech
Office Recording Device model	Laptop: - DELL (Latitude E5450) - LENOVO (20ARS0YB08)
No. of Speakers	30
No. of isolated words	501
No. of recorded sound	18,799 sound files (utterance)
Frequency	16000 Hz and 44100 Hz
Recording Channels	Mono and Stereo
Sound Files format	MS Wav (.wav)
Sampling Resolution	32-bit

3.4. Data Preprocessing

After collecting data as mentioned before, the following steps for data preprocessing.

1. Splitting individual words from sound files and saving each one as a new.wav file about the approximately 1-s duration for each one, using a model called "pydub" that can work with audio files, this library can play, split, merge, and edit.wav audio files.
2. After making chunks of the dataset facing many challenges, one of the challenges is appearing some sounds during recording like breathing in loud sounds "uhhh" the program treats as a separate word, should listen to each sound file carefully and discard these as an un-speech sound. Furthermore, some speakers make "umm" or "uh" sounds before or after reading words; in this case, this part has been removed from the entire speech, as a result, it does not affect the dataset.
3. During splitting sound into small chunks (separate words), these small files are read by a package for audio analysis like music generation and ASR. Improves building blocks necessary to create music information retrieval system. Mainly retrieves numerical Numpy array, which represents sound data. Moreover, sampling rate SR is the number of samples taken per second. By default, samples the file at a sampling rate of 22050 Hz, this sample rate could be overridden to any desired SR (8000, 11025, 16000, 22050, 32000, 44100, etc.) Hz. SR = number of samples per second. Taking from a continuous signal to make a discrete or digital signal, choosing a 16000 sample rate.
4. In Fig. 1, easily note that the speech part is ended in 1.2 s and mentioned before in challenge two the 0.2 s, after

TABLE 2: Samples of dataset

Sorani reading style	Latin reading style	Meaning in english
سفر	Sifr	Zero
یەک	Yek	One
دوو	Dû	Two
سێ	Sê	Three
چوار	Çwar	Four
پێنج	Pênç	Five
شەش	Şeş	Six
حەوت	Hewt	Seven
هەشت	Heşt	Eight
نۆ	No	Nine
دە	De	Ten
زانکۆ	Zanko	University
خوێندکار	Xwendkar	Student
یاریدەدەر	Yaridadar	Assistant
پزیشکی	Pzishky	Medicine
ئەنجومەن	Anjuman	Council
زانست	Zanst	Science

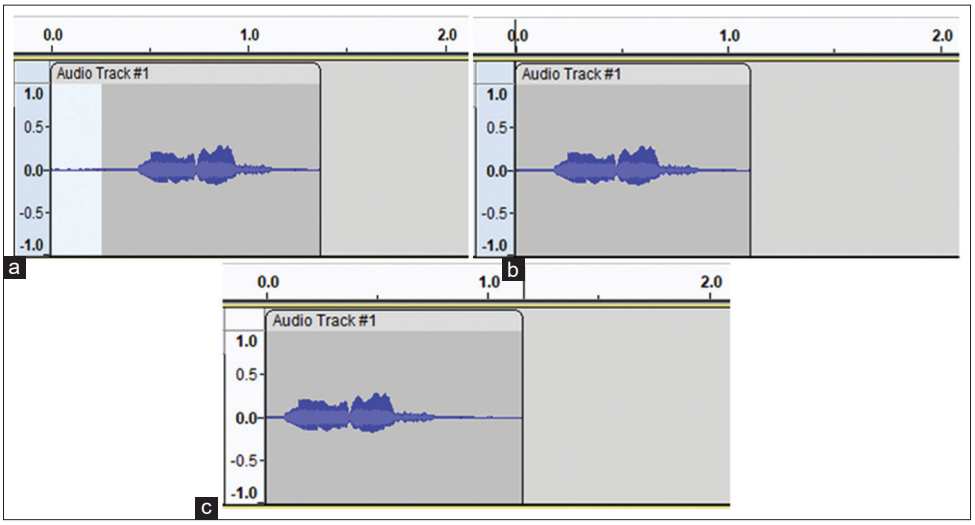

Fig. 1. The speech signal (a) with silence part, in the beginning, (b) with the removed silent part, and (c) with silence part in the end.

1.0 s will be removed and the speech will be unclear. In this case, should remove the silence part from the beginning of the speech, but if the silent part is after the speech, do nothing. The duration of the speech signal is 1.0 s, after that time, it does not matter and fixed it before in challenge two.

5. In some cases, when the word contains two or more sections like "هاوپێچ" which is mean "attach" the reader read it separately, the program treats it as a separate word which makes it mean less "پێچ" "هاو," fixed this problem manually by combining these sections.
6. In opposite to challenge four, in some cases, readers read two different words without any silence between them, and the program selected it as one word, also fixed it manually by separating them like "هەرێمی کوردستان" meaning "Kurdistan Region," as shown in Fig. 2.

1. The format of sound data retrieved from Viber was. m4a which is not supported by Audacity, should convert to a.wav file and also change its sampling rate from 48000 Hz to 16000Hz the size of the file reduced for instance a file size of 7.60 MB becomes 3.8 MB.
2. To prepare the data for fitting into a model, down sampling was applied to the recorded sound files, resulting in 8000 samples per word.

3.5. Data Augmentation (DA)

DA is the method of applying minor modifications to our original training dataset to produce new artificial training samples. There are many types of DA which are time wrapping, frequency masking, time masking, noise reduction, etc.

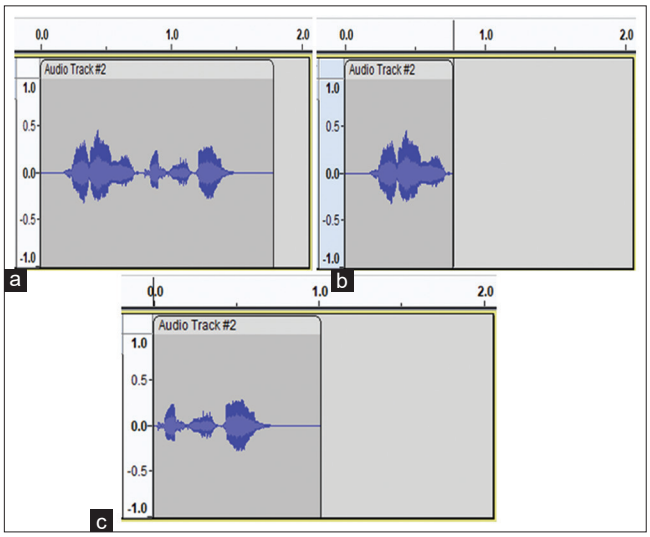

Fig. 2. Separate speech signal to individual words, (a) single chunk with two words, (b) separate first word, and (c) separate second word.

As in [23] used Additive White Gaussian Noise, pitch shifting, and stretching of the signal level. In the proposed work, since the number of speech utterance records in each class is relatively low, this study performs one type of audio DA, which is noise reduction.

4. METHODOLOGY

Both CNN and RNN have widely used in the speech recognition area and approved satisfactory results, for the Kurdish language using these models was challenging. This study proposed four different architecture models,

model generalizations, changing hyper parameters like batch size [24], and optimizers. Related to the study result concluded with CNN lower accuracy, then decided to go with a combination of CNN with RNN (GRU) to seek higher accuracy than the first model architecture.

4.1. CNN Model

The initial model architecture, the CNN model, was used to train and test the KSS dataset. The model comprises four CNN layers, as shown in Fig. 3, each of which is made up of three basic layers, the first of which is conv_layer. This is the first layer, and it is used to extract various features from the input data by performing mathematical operations between the input data and a filter of a specific size (8 × 8, 16 × 16, 32 × 32, 64 × 64) for each CNN layer. The second one A Pooling Layer is usually followed by a Convolutional Layer in most circumstances. This layer's major goal is to lower the size of the convolved feature map to reduce computational expenses. The third one over fitting happens when a model performs so well on training data that it hurts its performance when applied to new data. A dropout layer is used to solve this problem, in which a few neurons are removed from the neural network during the training process, resulting in a smaller model. After passing a dropout of 0.3, 30% of the nodes in the neural network are dropped out at random.

4.2. RNN (GRU)

The vanishing gradient problem affects RNN during back propagation. Gradients are values that are used to update the weights of a neural network. When a gradient reduces as it back propagates through time, this is known as the vanishing gradient problem. When a gradient value falls below a certain threshold, it no longer contributes much to learning. RNNs can forget what they have seen in longer sequences, because these layers do not learn, resulting in short-term memory. As a solution to short-term memory, LSTMs and GRUs were developed. They have inbuilt devices known as gates that can control the flow of data.

GRUs are a recurrent (RNN) gating technique first introduced in [25]. The GRU is similar to an (LSTM) with a forget gate, as shown in Fig. 4, but it has fewer parameters and lacks an output gate. Its performance in polyphonic music modeling, speech signal modeling, and natural language processing was found to be comparable to that of an LSTM. On certain smaller and less frequent datasets, GRUs have been proven to perform better [26].

4.3. Training Models

The ability of the model to adapt to new previously unseen data derived from the same distribution as the one used to generate the model is referred to as generalization. For instance, adding or removing layers to the current model leads to changing the accuracy of the system. Table 3 presents four different architecture models. After getting results from the Table 3, model (3) was chosen as the best model architecture. The detailed layers of the proposed architecture are shown in Fig. 5.

5. RESULTS AND DISCUSSION

In this section, the result of our experiments is presented in two different algorithms, the first is CNN, and the second is CNN+RNN (GRU), as explained in Table 3. As shown in Table 4 indicate that, each batch size got an accuracy. For the CNN model, architecture concluded the best batch size which is Batch size = 16.

To carry out a thorough investigation and achieve a fair comparison between the varied systems, the research uses the 10:90 and 20:80 approaches for testing and training, respectively, as a way of assessing the proposed systems. The results can then be averaged to compute a single estimation. This is particularly important when carrying out experiments with limited data sources, it is important to be clear that the point of this experiment was to ascertain how much data should reserve for testing. The result is presented in Tables 5 and 6 for the 10:90 splitting dataset and Table 7 for the 20:80 splitting dataset.

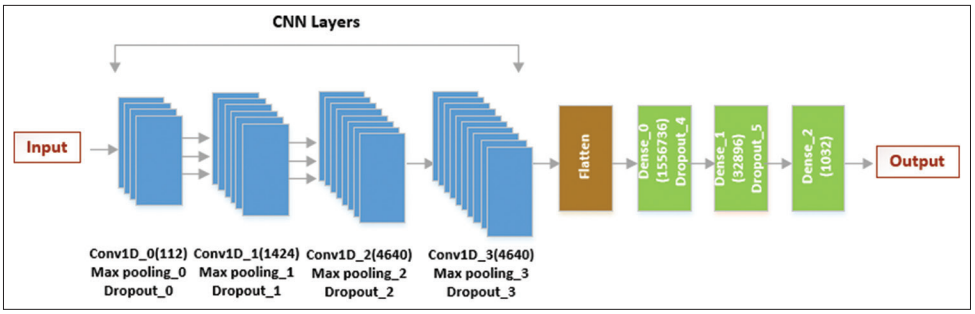

Fig. 3. The CNN architecture model.

After realizing combining three layers of CNN with three layers of RNN which is model, Architecture 3 shows a better result among other 4 architectures, as shown in Table 5, with %90 of the dataset for training the model and %10 for testing the model, after changing hyper parameters like batch size, also using SGD and Adam optimizer, the result shows in Tables 6 and 8.

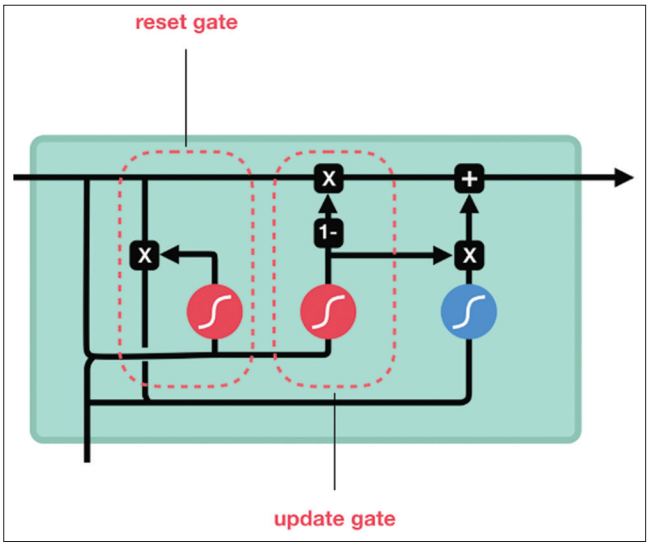

Fig. 4. Gated recurrent unit (Chung *et al*. 2014).

TABLE 3: Different types of model architecture with different layers

Layers	Model 1	Model 2	Model 3	Model 4
Conv_0 ✓	✓	✓	✓	✓
Maxpooling_0	✓	✓	✓	✓
Dropout_0	✓	✓	✓	✓
Conv_1	✓	✓	✓	✓
Maxpooling_1	✓	✓	✓	✓
Dropout_1	✓	✓	✓	✓
Conv_2	✓	✓	✓	×
Maxpooling_2	✓	✓	✓	×
Dropout_2	✓	✓	✓	×
Conv_3	✓	✓	×	×
Maxpooling_3	✓	✓	×	×
Dropout_3	✓	✓	×	×
flatten	✓	✓	×	×
Batch Normalization	×	×	✓	✓
GRU Bidirectional_0,1,2	×	×	✓	✓
Batch Normalization	×	×	✓	✓
Flatten	×	×	✓	✓
Dense_0	✓	✓	✓	✓
Dropout	✓	✓	✓	✓
Dense_1	✓	✓	✓	✓
Dropout	✓	✓	✓	✓
Dense_2	✓	✓	✓	✓
Dropout	✓	×	×	×

As discussed above indicate that both types of optimizer SGD and Adam could be used for speech recognition as they show a confident result, Adam optimizer reached the result that in batch size (64 and128), shows a better choice as it is accuracy reached (96% and 96%), respectively, which is higher than among batch size (8, 16, and 32) results (92%, 93%, and 72%). On the other hand, the SGD optimizer represents the result in different batch size values which are (16, 32, 64, and 128), but only in (32) reaches the result to (90%). This experiment discovered that the

TABLE 4: CNN model batch size and accuracy number

Batch size	Accuracy %
8	0.51661
16	0.61993
64	0.58672
77	0.47
99	0.40

TABLE 5: The accuracy for each model

Layers	Model 1	Model 2	Model 3	Model 4
Accuracy	0.61	0.88	0.92	0.87

TABLE 6: Using different batch size with ADAM optimizer

Optimizer	Batch size	Epochs	Accuracy %
Adam	8	31	0.92074
Adam	16	59	0.93738
Adam	32	58	0.7278
Adam	64	60	0.9601
Adam	128	69	0.96436

TABLE 7: The effect of splitting dataset to 20:80 on accuracy

Optimizer	Batch size	Epochs	Accuracy %
Adam	8	62	0.89734
Adam	16	59	0.94176
Adam	32	40	0.93697
Adam	64	49	0.94495
Adam	128	50	0.94441

TABLE 8: Using different batch size with SGD optimizer

Optimizer	Batch size	Epochs	Accuracy %
SGD	16	96	0.83989
SGD	32	16	0.90160
SGD	64	29	0.01489
SGD	128	31	0.01755

TABLE 9: Comparison with recent works related to proposed method, dataset, accuracy acheivments

Author	Proposed	Dataset	Acc.
Alkhateeb [12]	Arabic digit classification system using probabilistic neural network PNN.	450 Arabic spoken digits	93%
Arun *et al*. [18]	Malayalam speech recognition, The RNN was used.	5–10 solitary words	91%
Zerari *et al*. [21]	Used RNN for Arabic speech recognition.	Consists of 8800 tokens and TV command dataset: 10000 tokens for 10 TV commands	96%
Proposed System	Using CNN with RNN (GRU) for Kurdish word recognition	KSS Dataset that compose of 18799 sound files for 500 formal Kurdish words	96%

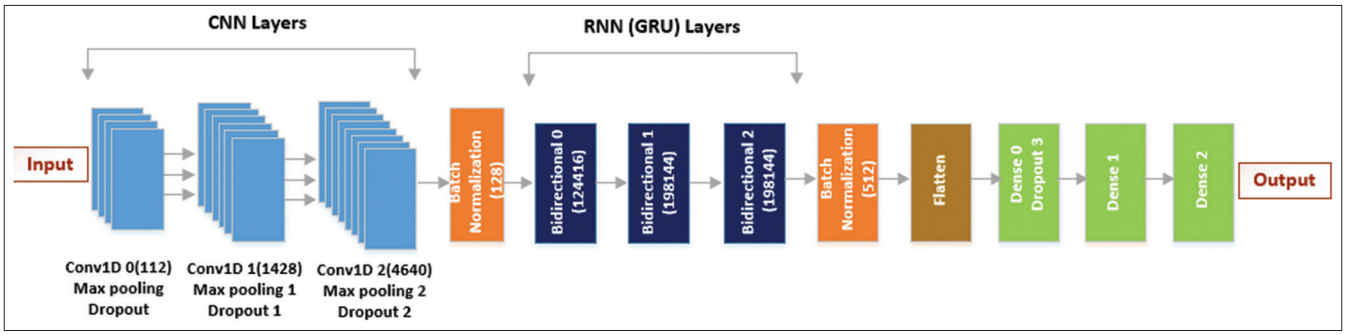

Fig. 5. The proposed architecture CNN_GRU model.

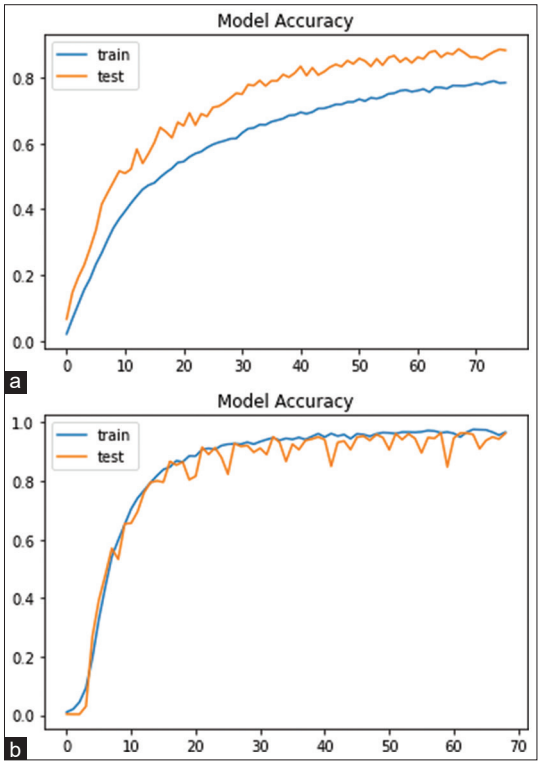

Fig. 6. (a) Accuracy for model CNN. (b) Accuracy for model CNN and RNN (GRU). Horizontal line indicates accuracy, while vertical line indicates number of Epochs.

Adam optimizer is more proper for speech recognition as getting a high accuracy in almost all the tests than the SGD optimizer. Fig. 6 shows the best accuracy for model CNN and RNN (GRU).

Now by changing the splitting dataset to 20:80 for testing: Training with batch size (8, 16, 32, 64, and 128), the results show in Table 7.

By comparison with recent works, the Table 9 indicates the comparison between those papers referenced with proposed method, its dataset and accuracy achieved.

6. CONCLUSION

In this paper, implemented speech recognition for the Kurdish language as well as created a KSS data set for this research purpose. The data set composed of 18799 sound files for 500 formal Kurdish words were read by 30 native Kurdish speakers. The research work designed different model architectures with different parameters using CNN and CNN+ RNN(GRU), the experimental findings indicate that the accuracy of the model increases when three layers of GRU are added to three layers of CNN. The accuracy of the CNN model reaches 61%, but after adding GRU, the accuracy increases dramatically to 96%, providing us with a clear vision for selecting the desired architecture. In the future, intend to improve the quality of Kurdish language materials, as well as utilize state of the art methods such as

metaheuristic optimizer with deep learning, to improve the performance.

REFERENCES

[1] E. Morris. "Automatic Speech Recognition for Low-Resource and Morphologically Complex Languages". *Thesis. Rochester Institute of Technology*, 2021.

[2] S. Ruan, J. O. Wobbrock, K. Liou, A. Ng and J. A. Landay. "Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones". *Journal Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies Archive*, vol. 1, no.4, pp. 1-23, 2017.

[3] M. Assefi, M. Wittie and A, Knight. "Impact of network performance on cloud speech recognition". In: *Proceedings of the 24th International Conference*, pp. 1, 2015.

[4] M. Asseffi, G. Liu, M. P. Wittie and C. Izurieta. "*An Experimental Evaluation of Apple Siri and Google Speech Recognition*". ISCA SEDE Montana State University, Bozeman, 2015.

[5] A. Ganj and F. Shenava. "2-Persian continuous speech recognition software". In: *The First Workshop on Persian Language and Computer*. The 9th Iranian Electrical Engineering Conference, Iran, 2004.

[6] F. A. Ganj, S. A. Seyedsalehi, M. Bijankhan, H. Sameti, S. Zadegan and J. Shenava. "1-Persian continuous speech recognition system". In: *The 9th Iranian Electrical Engineering Conference*, 2000.

[7] A. Qader and H. Hassani. "*Kurdish (Sorani) Speech to Text: Presenting an Experimental Dataset*". arXiv: 1911.13087v1, 2019.

[8] R. Yaseen and H. Hassani. "Kurdish Optical Character Recognition". *UKH Journal of Science and Engineering*, vol. 2, pp. 18-27, 2018.

[9] R. D. Zarro and M. A. Anwer. "Recognition-based online Kurdish character recognition using hidden Markov model and harmony search Eng." *Engineering Science and Technology an International Journal*, vol. 20, no. 2, pp. 783-794, 2017.

[10] A. T. Tofiq and J. A. Hussain. "Kurdish Text Segmentation using projection-based approaches". *UHD Journal of Science and Technology*, vol. 5, no. 1, pp. 56-65, 2021.

[11] H. Veisi, H. Hosseini, M. Amini, W. Fathy and A. Mahmudi. "*Jira: A Kurdish Speech Recognition System Designing and Building Speech Corpus and Pronunciation Lexicon*". ArXiv abs/2102. 07412, 2021.

[12] A. Alkhateeb. "Wavelet LPC with neural network for spoken arabic digits recognition system". *Jordan Journal of Applied Science*, vol. 4, pp. 1248-1255, 2014.

[13] N. Turab, K. Khatatneh and A. Odeh. "A novel arabic speech recognition method using neural networks and gaussian filtering". *IJEECS International Journal of Electrical, Electronics and Computer Systems*, vol. 19, pp. 1-5, 2014.

[14] S. Malekzadeh, M. H. Gholizadeh and S. N. Razavi. "*Persian Phonemes Recognition Using PPNet*". arXiv preprint arXiv: 1812.08600, 2018.

[15] H. Veisi and A. Haji Mani. "Persian speech recognition using long short-term memory". In: *The 21st National Conference of the Computer Society of Iran*. University of Tehran, Iran, 2015.

[16] A. Graves, A. R. Mohamed and G. Hinton. "Speech recognition with deep recurrent neural networks". In: *ICASSP Conference*. Institute of Electrical and Electronics Engineers, Piscataway, 2013.

[17] A. R. Mohamed, G. Dahl and G. Hinton. "Deep belief networks for phone recognition". In: *Nips Workshop on Deep Learning for Speech Recognition and Related Applications*. . IJCA Proceedings on National Conference, USA, 2009.

[18] H. P. Arun, J. Kunjumon, R. Sambhunath and A. S. Ansalem. "Malayalam speech to text conversion using deep learning". *IOSR Journal of Engineering* (*IOSRJEN*), vol. 11, no. 7, pp. 24-30, 2021.

[19] M. M. H. Nahid, B. Purkaystha and M. S. Islam. "Bengali speech recognition: A double layered LSTM-RNN approach". In: *Procceding 20th Institute of Communication Culture Information and Technology*, pp. 1-6, 2017.

[20] M. Ravanelli, P. H. Brakel, M. Omologo and Y. Bengio. "Light gated recurrent units for speech recognition". *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, pp. 92-102, 2018.

[21] N. Zerari, S. Abdelhamid, H. Bouzgou and C. Raymond. "Bidirectional deep architecture for Arabic speech recognition". *Open Computer Science*, vol. 9, pp. 92-102, 2019.

[22] R. Ahmed, S. Islam, A. K. M. Muzahidul Islam and S. Shatabda1. "An Ensemble 1D-CNN-LSTM-GRU Model with Data Augmentation for Speech Emotion Recognition". arXiv: 2112.05666, 2021.

[23] C. Huang, G. Chen, H. Yu, Y. Bao and L. Zhao. "Speech emotion recognition under white noise". *Archives of Acoustics*, vol. 38. pp. 457-463, 2013

[24] I. Kandel and M. Castelli. "The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset". *ICT Express*, vol. 6, no. 4, pp. 312-315, 2020.

[25] K. Cho, B. V. Merrienboer, D. Bahdanau and Y. Bengio. "*On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*". arXiv: 1409.1259v2, 2014.

[26] J. Chung, C. Gulcehre, K. Cho and Y. Bengio. "*Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*". arXiv: 1412.3555v1, 2014.