P- ISSN 1991-8941 E-ISSN 2706-6703 *Journal of University of Anbar for Pure Science (JUAPS)* Open Access
2022,16, (2) :108-118

Central Kurdish Automatic Speech Recognition using Deep Learning

Abdulhady Abas Abdullah[1*], Hadi Veisi[2]

[1] Computer Science Department-Faculty of Science - Soran University, Soran, Erbil, Kurdistan, Iraq aaa160h@cs.soran.edu.iq

[2] University of Tehran (Visitor at Soran University), Faculty of New Sciences and Technologies - h.veisi@ut.ac.ir, hadi.veisi@visitors.soran.edu.iq

ARTICLE INFO
Received: 28 /07 /2022
Accepted: 28 / 08 /2022
Available online: 23/12/2022

DOI: 10.37652/juaps.2022.176500

Keywords:

Automatic Speech Recognition;
Central Kurdish Language;
Deep Learning;
Transfer Learning.

ABSTRACT

Automatic Speech Recognition (ASR) as an interesting field of speech processing, is nowadays utilized in real applications which are implemented using various techniques. Amongst them, the artificial neural network is the most popular one. Increasing the performance and making these systems robust to noise are among the current challenges. This paper addresses the development of an ASR system for the Central Kurdish language (CKB) using a transfer learning of Deep Neural Networks (DNN). The combination of Mel-Frequency Cepstral Coefficients (MFCCs) for extracting features of speech signals, Long Short-Term Memory (LSTM) with Connectionist Temporal Classification (CTC) output layer is used to create an Acoustic Model (AM) on the AsoSoft CKB speech dataset. Also, we have used the N-gram language model on the collected large text dataset which includes about 300 million tokens. The text corpus is also used to extract a dynamic lexicon model that contains over 2.5 million CKB words. The obtained results show that the use of a DNN improves the results compared to classical statistics modules. The proposed method achieves a 0.22%-word error rate by combining transfer learning and language model adaptation. This result is superior to the best-reported result for the CKB.

1. Introduction:

Speech is a more common and primary approach to communication between humans. This communication between humans and machines was referee to by way of the human-computer boundary. The method of translating a voice signal into a sequence of words using a technique such as a computer program is known as Automatic Speech Recognition (ASR). The ASR field aims to build processes and systems for using speech to communicate with machines [1]. In addition to ASR, there are various applications for speech processing such as text-to-speech [28], detection of speech signals [2], speaker recognition and detection of speaker identity [3], speech emotion recognition [4], health recognition, and patient's health status [5], language recognition, spoken language knowledge [6], accent and dialect recognition, and gender recognition [7]. Nowadays, automatic speech processing has also developed into the newest and fastest system between human-machine interfaces [1].

———— *Corresponding author at Computer Science Department-Faculty of Science - Soran University, Soran, Erbil, Kurdistan, Iraq;ORCID:https://orcid.org/0000-0000-00000-;Tel:
E-mail address: aaa160h@cs.soran.edu.iq

ASR was an important study subject for many years and aims to enable realistic human-machine interaction. Just before the turn of the century, Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), Mel-Frequency Cepstral Coefficients (MFCCs) and derivatives, N-gram Language Models (LMs), discriminative training, and a few adaptation strategies were primarily created throughout the process [8].

Over the past ten years, the topic of Deep Learning (DL) or Deep Neural Networks (DNN) has gained popularity as a fresh Machine Learning (ML) area [9]. DL consists of many ML algorithms that feed input data as many-layered models. Typically, these systems were neural networks with several levels of non-linear processes. ML algorithms try to derive basic features and knowledge of DNN to learn from them [10]. Especially reviewing speech signal processing, we found that these techniques overcome the classical techniques of HMM and GMM to process speech signals [2].

Implementing an ASR system in real-time environments is difficult, especially for languages with low resources, background noise, and poor sound quality in the speech signal. Although ASR systems are

been developed for many languages such as English [15], Japanese [18], Russian [26], Turkish [19], Persian [20], and Arabic [21], however, the Kurdish language is not among them. Kurdish, as an Indo-European language, is categorized into three main branches, i.e., Central Kurdish (CKB, i.e., Sorani), North Kurdish (Kirmanji), and South Kurdish, and is spoken by more than 30 million people. In addition to the mentioned general challenges, another problem is that there is no standard spelling of words in CKB, which is one of the obstacles in the field of Natural Language Processing (NLP) in this language. As well, Kurdish is one of these languages with low resources. Limited studies are been conducted on Kurdish ASR before, but on a small scale, such as recognition digit [11], isolated spoken word recognition [12], or Kurdish spoken letter recognition [13]. Only one study on developing a large vocabulary speech recognition system for the Kurdish language was done by the AsoSoft team which 43 hours of Kurdish corpus have crated [14].

In this study, to overcome the mentioned challenges and improve the quality of the acoustic model, the Long-Term Short-Term Memory (LSTM) technique is applied to the CKB. In addition, we are taken advantage of N-gram language modeling for the output layer part of the network to enhance word accuracy performance. Also, adapting our model to new words and recognizing them by using a dynamic lexicon model which automatically extracts and collects a text corpus. Our lexicon model is over 2.5 million CKB words. We also used the language model to standardize the Kurdish spelling, extracting the largest spelling frequency of words in the text corpus and then converting the various spellings of the words in the text corpus to the most common spelling type. Therefore, the main contributions of this work are:

-Apply an end-to-end DNN to improve the performance Acoustic Model (AM) in real-time and different speaker environments.

-Collecting a large-scale text corpus for training the language model and extracting a large vocabulary lexicon.

-Utilize dynamic lexicon which consists of more than 2.5 million CKB words.

The structure of this paper is as follows: In Section 2, the related works with speech recognition technology are described. The proposed method and the architecture of the applied model for the CKB language ASR system are explained in Section 3. In Section 4, the result and discussion are shown. This paper is concluded in the final section.

2. RELATED WORK

This section reviews recent related works using DL techniques for ASR.

In [15], Recurrent Neural Networks (RNN) and LSTM techniques are implemented for a state-of-the-art ASR system; also, Wall Street Journal, Switchboard, and Fisher datasets are used to create the training model. The authors utilized Deep Speech [9] and achieved a 6.56% of Word Error Rate (WER) for clean speech, 19.06% of WER for noisy, and 11.85% for combined clean and noisy speech. In [16] convolutional neural network (CNN) used to train the ASR model, the authors achieved a frame error rate of 22.1%, which is quite near to the state-of-the-art. The encoded phone sequence has a 29.4% error rate. Another research implemented the general approach, including RNN and CTC, and used Noisy Dev Database as the trained dataset. The findings support and demonstrate the use of end-to-end DL techniques in speech recognition in various situations [9]. Moreover, [17] proposed a DL speech processing and recognition technique. The authors utilized general approaches such as CNN and Artificial Neural Network (ANN) and the trained dataset obtained in Google's dataset. Even though only a few hours of transcribed data were available, the model's accuracy was 66.22%.

For the first time, Farsdat Persian audio dataset was utilized to develop an AM using a combination of Deep Belief Network (DBN) for feature extraction in voice signals and "Deep Bidirectional Long Short-Term Memory (DBLSTM)" with a CTC SoftMax output-layer. Compared to the Kaldi-DNN and HMM, the findings show that employing DBLSTM with features derived from the DBN improves Persian phoneme identification accuracy [20]. This study investigates the advanced "end-to-end" DL technique for robust Arabic diacritical construction of ASR. For this purpose, each CTC, LSTM, MFCC, and CNN was used to train the model. SASSC dataset was used for train and testing. As a result, the WER with traditional oppressions decreased by 5.24% and 2.62% [20].

The Jira ASR system is introduced as the first "large vocabulary speech recognition system (LVSR)" for the CKB language. Several conventional methods are employed to create the acoustic model, including HMM-based models and SGMM approaches. Regarding speech corpus, the researchers formulated a

phrase collection di-phone ratio that closely reflects the CKB. 576 people spoke the intended word for a total of 43.68 hours in both a controlled environment using a noise-free microphone (aka AsoSoft Speech-Office) and a social network setting using a mobile phone (aka AsoSoft Speech-Crowdsourcing). The SGMM acoustic model achieves the most significant results, with an average word mistake rate of 13.9 % (on various document themes) and 4.9 % for the overall subject [14].

The summary related literature is presented in Table 1. This table shows that in the last year, most languages are used end-to-end deep learning techniques, including RNN, LSTM for acoustic model training and CTC for the output layer, and MFCC for feature extraction. The accuracy and speed of the models are improved significantly compared to older techniques. As a result, this review showed that deep learning techniques are very suitable for CKB according to the results obtained for other languages.

Table 1: Summary of review on deep learning techniques for speech recognition.

No.	Language	Dataset	Technique(s)	Performance	Reference
1.	English	Wall Street Journal, Switchboard and Fisher	LSTM and MFCC	Accuracy = 11.85 %	[15]
2.	English	TIMIT	CNN, CTC, and MFCC	WER= 22.1%.	[16]
3.	English	Noisy Dev	RNN, CTC, and MFCC	0.1-0.3% WER	[9]
4.	English	speech obtained in Google's dataset	CNN, CTC, and MFCC	Accuracy = 66.22%	[17]
5.	Russian	SPIIRAS	LSTM, CTC, and MFCC	WER= 27.83%	[26]
6.	Japanese	A Japanese audio-visual dataset	CNN, CTC, and MFCC	accuracy = 90%	[18]
7.	Turkish	Turkish speech dataset	LSTM, CTC, and MFCC	WER= 14%	[19]
8.	Persian	The first database used	DBN, Deep Bidirectional-LSTM, and MFCC	Accuracy = 8.1 %	[20]
9.	Arabic	1200 h corpus	CNN-LSTM, CTC, and MFCC	WER= 33.72%	[21]
10.	Kurdish	Asosoft dataset	HMM-GMM, DNN, and MFCC	WER= 25 %	[14]

3. PROPOSED METHOD

As shown in Figure 1, a typical ASR system has three models, i.e., AM, LM, and lexical model, which are prepared during the training/development phase. In addition to the model estimation methods, these systems use two main components, i.e., signal-processing and feature-extraction, and hypothesis search (decoder). The audio signal is sent into the signal processing and feature extraction component, which improves speech by eliminating noises and channel distortions, translates the signal from time to frequency domain, and generates prominent feature vectors appropriate for the following acoustic models. The acoustic model combines acoustics and phonetics information, uses the features extracted by the feature extraction component as input, and provides an AM-score for the variable-length feature sequence. In the final step, the output obtained from the AM model is quality enhanced by LM and CTC in the decode section of the network. The components are explained in detail in the following sections.

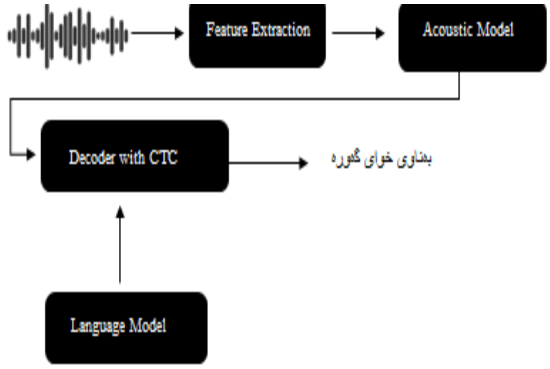

Figure 1. The ASR structure for this paper.

3.1 Corpus and Lexicon

As shown in Figure 1, to build an ASR system, we are to use two corpora: a speech corpus for training the AM and a text corpus to create an LM and a lexicon. We used the AsoSoft speech dataset [14] to train the AM of the Kurdish speech recognition system. The train speech corpus has a duration of 43.68 hours. Table 2 provides a general overview of the training dataset. The ASR system converts phonemes from the acoustic model into character chains, which are recognized as independent words. Converting the character chains into words requires a lexicon model that automatically extracts and collects a text corpus. Our model lexicon is over 2.5 million CKB words. The following is a more detailed description of these subsets. We also collected large text data from textbooks, websites, newspapers, journals, and magazines to train the language model. The total text corpus is 300 million tokens, making it the largest text corpus in CKB for training NLP models.

Table 2: AsoSoft Speech Corpus (train set) (Veisi et al., 2021)

	Number. of Speakers	Number. of Utterances	Durations (Hr)
Speech in Office Room	**60.0**	**31,075**	**31.520**
Crowdsourcing Speech signal	**516.0**	**11,519**	**12.160**
Total	**576.0**	**42,594**	**43.6800**

3.2 Extracting Features using MFCC

The feature extraction stage determines the collection of phrase features that are an acoustic relationship with speech signals, and acoustic waveform processing was used to create these parameters. These qualities are often referred to as features. The primary goal of a feature extractor is to maintain useful data while discarding irrelevant data [21]. A block diagram of the MFCC architecture is presented in Figure 2. Table 3 also shows the configuration parameters for the application in the speech corpus.

Table 3. MFCC parameters

Parameter	Value
Frame size	**256 samples**
Frame overlap	**128 samples**
Pre-emphasis coefficient (α)	**-0.95**
Number of triangular band-pass filters	**20**
Number of MFCC coefficients	**13 with energy**

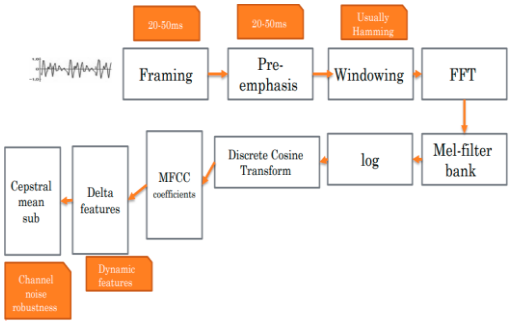

Figure 2. MFCC from the audio recording signals [22].

3.3 Acoustic Modelling

This method represents the association between text and its sound recognitions by a succession of stochastic generative acoustic models. For each word or sub-word unit, the functionality of an automated speech recognition system may be characterized as the extraction of several speech characteristics from the acoustic speech signal [18]. One of the most common techniques today is the RNN technique, which in recent years has completely replaced the older statistical techniques such as HMM-GMM. This great advance in the use of RNN is since it solves the problems of ASR very well, in real-time and in various noisy environments. It also provides great support for languages with limited resources by making technical recommendations for transfer learning and finetuning. In this work, we used LSTM, a type of RNN recently very popular in ASR, which improves the performance of the model as described in detail in Section 4.

Model Architecture DeepSpeech[9] is a bidirectional end-to-end LSTM deep learning created exclusively for ASR. It has six layers, then fully connected, LSTM, and a SoftMax layer. An input layer including Mel Spectrograms as the initial input layer. The first three layers were fully connected through a ReLU activation function, while the four layers are made up of an LSTM unit; before the five layers are fully connected again, ReLU is an activation function. The final layer, a fully connected layer with a SoftMax activation function with normalizing, outputs probability for every letter inside the language's letter. A CTC loss function is also calculated using the letter probability [42]. In terms of the CTC loss, the

parameters of a model were optimized to use the Adam approach. A general Structure of the training in LSTM is shown in Figure 3.

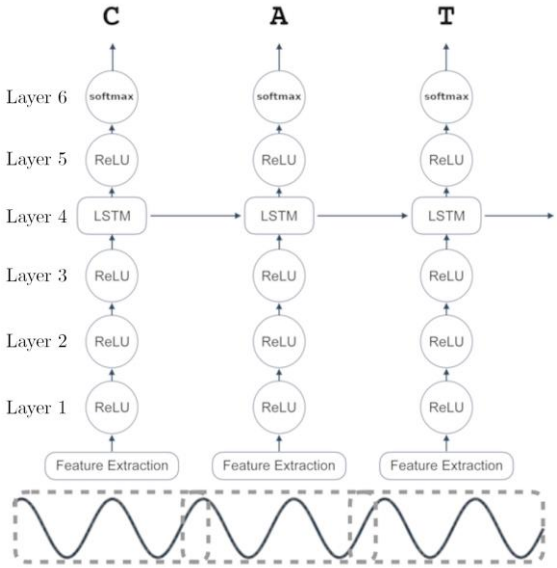

Figure 3. The LSTM (Deepspeech2) Structure [9]

3.4 Connectionist Temporal Classification (CTC)

One of the drawbacks of employing DL in ASR is that each frame of the incoming signal is labeled independently by the neural network. This means the network flags each frame individually rather than supplying the phonetic sequencing based on the input signal. As a result, to acquire the phoneme sequence corresponding to an audio signal, post-processing algorithms should be utilized to extract a phonetic sequence matching input data from every frame's outputs. The standard output layer of a neural network can be replaced with a CTC layer as one solution to the crisis [23]. In CTC, examples of data elements are shown in Figure 4. Only the output-layer neurons' delta computations alter in the CTC approach, and the network structure might be bidirectional, unidirectional, deep, or profound. In temporal classification tasks, RNNs' CTC output layer is employed. In other words, this approach is used to identify the sequence in which the mappings between both the source & output sequences are unclear. As a consequence, no further processing is necessary to get the label sequences from the neural network outputs using this approach. Instead of identifying each frame, the network generates a phoneme sequence for the entire signal. Forward-backward and decoding are two separate algorithms that make up the CTC algorithm.

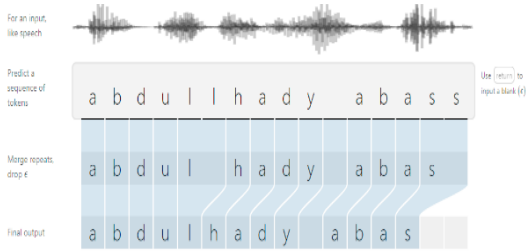

Figure 4. Row data of CTC Structure

3.5 Language Model

The language model, which limits the decoder's search space in an ASR system, is yet another useful source of data for detecting ending words. the n-gram language model is one of the most widely used tools in language processing. The n-gram is a popular technique because the model is based on counting the probability of observing each possible bi-gram. This is an efficient way to make use of the data, especially when you do not have a lot of text to train from. N-gram models can easily beat neural network models on small datasets. CKB is one of the languages with few textual data sources; on this basis, we used this technique.

We examine two types of language models: a 3-gram model and a 4-gram model, both of which are trained on a 300 million token corpus. A text corpus of 300 million tokens dataset was mostly acquired from books, periodicals, newspapers, and internet websites. Text normalization is an essential step in improving the quality of language models, and it's even more critical in the case of Kurdish because the authors and publishers use various encoding systems and orthographic rules. This paper discusses the challenges and some solutions for text normalization [30].

3. 6 Implementation Environment and Hyper-Parameters

In this section, we discuss training for the language model and acoustic model.

For the exercises, all models used a Linux machine with an Intel I9 9900K @ 2.10GHz CPU, 500 GB SSD memory, and an RTX 3080 10GB Gainward Ghost GPU with 10 GB memory.

First, we used a 300 million token text corpus to train the language model using N-gram statistical technique. For CKB, we ran several tests with different parameters to acquire the best results; in the beginning, we trained the amount of top_k=500000 words at 96.3040% of all words, we also used the 4-gram, and we used max_arpa_memory=85%. As a result, our dynamically obtained model lexicon consisted of 500,000 words that covered 96% of the words in our

text corpus. Our lexicon was as follows: Your most common word "باشه" occurred 12059567 times, the least common word in your top-k is "مەشقەکیان" with one time, the first word with two occurrences is "پاراسایکۆلۆژی."The second evaluation was the amount of top_k=2500000 words at 99.6090% of all words; we also used the 3-gram and max_arpa_memory=85%. As a result, our dynamically obtained model lexicon consisted of 2500000 words that covered 99.6% of the words in our text corpus in the text file has 299207607 words in total it has 3939212 unique words.

As a result, the total time it took for both models to train the ARPA file, building lm, binary, vocabulary, and KenLM-scorer file was one hour. Also, the performance of the second model was much better than the first model. Furthermore, we used the AsoSoft speech corpus to train and test the acoustic model consisting of 43 hrs. of data. First, we divided the corpus into training, validation, and testing, 75% for training, 15 % for testing, and 10 % for validation. The total number of sentences for the train set is 30,281, which is 35 hours of data, the number of validations set sentences is 765, which is 3 hours, and the test set is 4,934 sentences, which is 5 hours for the train acoustic model. Here we focus on training our best model; for the train acoustic model, we used the LSTM technique and created many other trains with different parameters, including n_hidden= 2048-layer width to use when initializing layers, learning_rate= 0.00001 Adam optimizer learning rate, drop rate = 0.40 drop rate for presentation layers, UTF-8 mode enabled for CKB language. The network output UTF-8 sequences alphabets directly rather than alphabet mapping to phonemes. The model allows an early stopping mechanism for the dataset validation. We also used 10 and 100 epochs to train the network, and each epoch required 40m for a total of 66h of training time. Also, we use transfer learning techniques to improve the model's performance based on the pre-trained English language model. Kurdish has a different alphabet from English; the Kurdish alphabet is utf8, so we cannot use the finetuning technique. In this case, we use the transfer-learning technique for the new alphabet. ASR implementation of transfer-learning, all removed layers must be contiguous and include the output layer. The flag to control the number of layers you remove from the source model is --drop_source_layers. This flag accepts an integer from 1 to 5, specifying how many layers to remove from the pre-trained model. This work has removed only one layer for CKB, the output layer because the Kurdish alphabet is different from our pre-trained model.

4. RESULTS AND DISCUSSION

This section presents the proposed method's results and comparison with the Jira paper HMM-GMM

4.1 ASR Evaluation Criteria

ASR systems' performance is usually measured in terms of accuracy and speed. Accuracy is generally measured in terms of performance accuracy, usually expressed as a word error rate (WER), whereas speed is described as a real-time factor. Single Word Error Rate (SWER) and Command Success Rate are two other accuracy measures (CSR). Word Error Rate (WER) and Word Recognition Rate (WRR) are used to evaluate the speech recognizer's performance. Several types of word errors include insertions, substitutions, and deletions. Finally, the following equations are used to calculate the word error rate and word recognition rate.

$$Word\ Error\ Rate(\%) = \frac{Insertion(1) + Substitution(s) + Deletion(D)}{No.of\ Reference\ Words(N)} * 100 \quad (1)$$

$$\text{Word Recognition Rate (WRR)} = 1 - \text{WER} = \frac{N - S - D - 1}{N} \quad (2)$$

The WER measure is explained in the following examples:
Best WET: means the currency sentence for the test is equal to the prediction sentence in the model:
WER: 0.000000, CER: 0.000000, loss: 0.00
file: F01146030.wav
test: " گشتپرسییە سەبارەت بە هەموارکردنەوەی دەستوورە "
prediction:" گشتپرسییە سەبارەت بە هەموارکردنەوە دەستوورە "

Median WER: The second example is WER smaller than 50%:
WER: 0.200000, CER: 0.056604, loss: 28.824883
file: M01143044.wav
test: "تیشک گەردیلەکان لە یەک بدەین ورد دەبن "
prediction: "تیشک گەردیلەکان لە یەک بدەین گردەبن "

Worst WER: Third example WER greater than 50%:
WER: 0.857143, CER: 0.470588, loss: 15.005235
file: M01142003.wav
test: "داوای بەخشش و یارمەتی لە خودا بکەن"
prediction: "داوای بەیارمەتیدان"

4.2 Experimental and Evaluating Results

In this section, WER is reported for 12 models. All models are trained, validated, and tested on the

same speech corpus. The results in Table 4 show the WER for all the models of the ASR. All the results in Table 4 were trained on the 4-gram language model, a lexicon model of 500,000 words, and the audio models were trained with the same parameters. The parameters are listed in Table 4. The test serves as our absolute baseline, in which the entire six layers of the CTC model were trained from scratch without using any transfer from English. Table 4 concludes that the scratch train model achieved 92 % WER.

In the second experiment, in which the five layers of the ASR network were created from scratch, only one layer of the English model was copied, the fully connected layer of the network, resulting in a 29% reduction compared to the first model, and a reduction in WER 63%.

In the third experiment, in which the four layers of the ASR network were created from scratch, two layers of the English model were copied to fully connect to the first network, resulting in a 40% reduction compared to the first model and a 52% WER reduction.

In the fourth experiment, in which the three layers of the ASR network were created from scratch, three layers of the English model were copied of fully connected (2 of ReLU) and LSTM layers in-network, resulting in a 43% reduction compared to the first model, and a 49% WER reduction.

In the fifth experiment, in which the two layers of the ASR network were created from scratch, four layers of the English model were copied of fully connect (2 of ReLU), LSTM, and fully connect (ReLU) layers in-network, resulting in a 54% reduction compared to the first model, and a 38% WER reduction. In the sixth experiment, in which one layer of the ASR network was created from scratch, this output layer softmax and CTC, five layers of the English model were copied of fully connect (2 layers of ReLU), LSTM, and fully connect (2-layer of ReLU) layers in-network, resulting in a 68% reduction compared to the first model. A 31% WER reduction in the model's structure is explained in Figure 5.

Furthermore, the removal and retention of layers are a significant effect on the accuracy of the model, because there is a very low speech corpus available for training. The copied layers improve the quality of the model because it has already been trained on large data for English. It learns recognition features very well. We only remove the English model alphabet and retrain it on the Kurdish alphabet. Thereby, the quality of the obtained model is greatly enhanced, as Tables 4 and 5 show the results.

As a result of these experiments in Figure 6, we found that DL techniques require a lot of data to train from scratch, as we saw for CKB for 43 hours of data gave us very debilitated results with an accuracy of 8%, that is, only eight words per 100 words correctly recognized. After applying transfer learning and the pre-training model, the WER is further reduced by 31%. The second result shown in Table 5 shows the WER for all ASR models. All the results in Table 5 are trained on the 3-Gram language model, which contains a lexicon model of 2.5 million words and trained to acoustic models with the same parameters and parameters listed in Table 5.

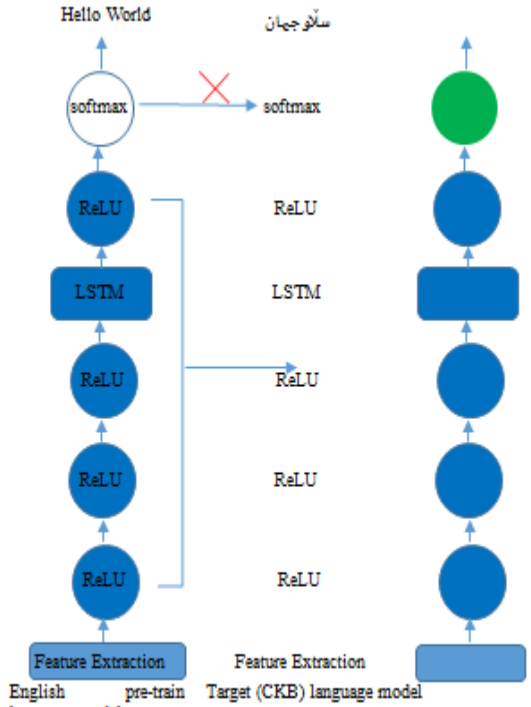

Figure 5. transfer learning coded five layers, and one layer of the ASR
network was created from scratch.

The results obtained in Table 5 compared to Table 4 were to change some of these parameters in the language model section, increase the lexicon data from 500,000 to 2.5 million words, and reduce the model from 4 grams to 3 grams. The acoustic model part was just changing the learning rate parameter from 0.0001 to 0.00001.

As a result, all results in Figure 7 are significantly smaller according to WER compared to Figure 6.

4.3 Comparison of the HMM-GMM with the LSTM model

As a result, the proposed model performs much better extraction than the previous model for CKB in

continuous time. Our model also performs much higher in terms of recognizing new words and sentences than the Jira model because we used dynamic lexicons. The language model trained on 300 million tokens is much higher than the previous model; the best results are shown in Table 6. This work reduced WER to %22. in continuous time, the speech signal increases in uncertainty, which is an obstacle to speech recognition. In addition, a DL technique improved the quality and enhanced the model's performance.

Table 6 Comparison of the Jira model with our models

Method	WER
Jira HMM [14]	**.25**
LSTM + 4-gram	**.31**
LSTM + 3-gram	**.22**

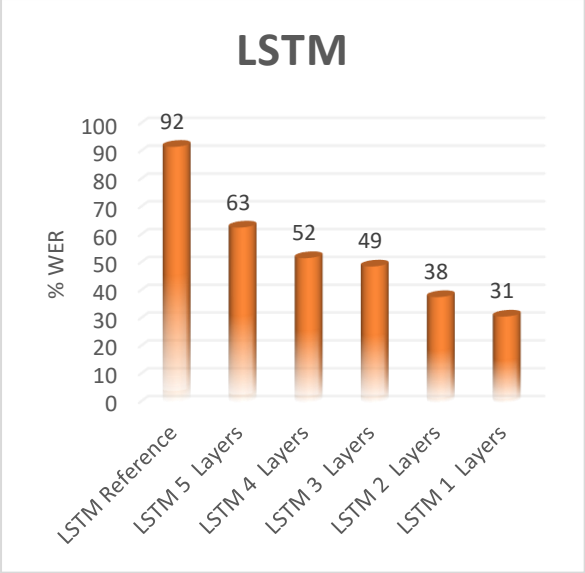

Figure 6. WER for language model 4-gram and acoustic model

Table 4: The best CER and WER for ASR Kurdish Language on the test set by using 4-gram

Method	Reference	*0 Frozen Layers*	*1 Frozen Layer*	*2 Frozen Layers*	*3 Frozen Layers*	*4 Frozen Layers*
N-gram	4	4	4	4	4	4
Top_k	500000	500000	500000	500000	500000	500000
n_hidden	2048	2048	2048	2048	2048	2048
Learning rate	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Dropout rate	0.4	0.4	0.4	0.4	0.4	0.4
Epoch	100	10	10	10	10	10
Batch size	64	64	64	64	64	64
WER	0.92	0.31	0.38	0.49	0.52	0.63
CER	0.74	0.45	0.29	0.11	0.09	0.26

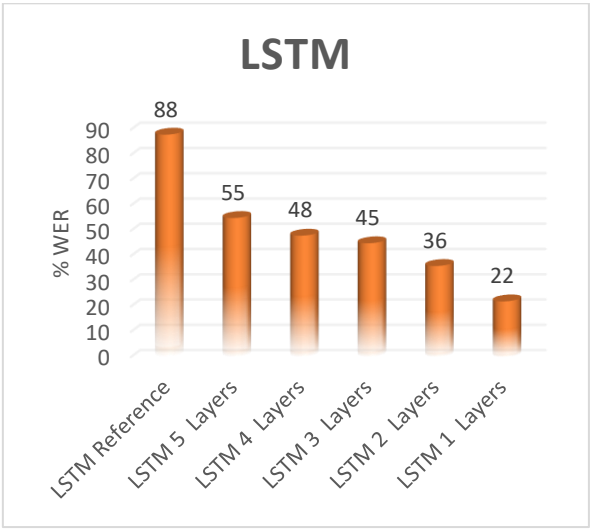

Figure 7. WER for language model 3-gram and acoustic model

Table 5: The best CER and WER for ASR Kurdish Language on the test set by using 3-gram

Method	Reference	*0 Frozen Layers*
N-gram	3	3
Top_k	2500000	2500000
n_hidden	2048	2048
Learning rate	0.00001	0.00001
Dropout rate	0.4	0.4
Epoch	100	10
Batch size	64	64
WER	0.88	0.22
CER	0.74	0.45

	1 Frozen Layer	2 Frozen Layers	3 Frozen Layers	4 Frozen Layers
	3	3	3	3
	2500000	2500000	2500000	2500000
	2048	2048	2048	2048
	0.00001	0.00001	0.00001	0.00001
	0.4	0.4	0.4	0.4
	10	10	10	10
	64	64	64	64
	0.36	0.45	0.48	0.55
	0.29	0.11	0.09	0.26

5. CONCLUSION

In this study, we described our efforts to build and implement the first ASR for CKB using DL and the constraints that go along with it, such as a speech corpus and a language model without a static pronunciation lexicon. DL has become a very important topic for researchers over the past few years. Without using a phonetic lexicon, we propose using a combination of transfer learning and language model adaptation to customize generic models to the unique properties of our data. Furthermore, research examining the efficacy of integrating transfer learning with language model adaptation for the low resource (Kurdish Language) ASR is limited. For low-resource data, the outcomes of transfer learning and language model adaptation are presented. As a result, the proposed RNN and LSTM algorithms also produce the lowest WER for the AsoSoft train set (i.e., 0.22 %). The addition of a language model improved systems; however, because of the particular vocabulary of our dataset, the language models trained, in part or whole, with inmate-like speech, offered the best results. The ASR system exhibited the most improvement after incorporating a language model, with WER dropping from 0.22 %. We intended to enhance the quantity the language skills for such a Kurdish language from 43 hrs. The results showed that the proposed techniques improved the quality and performance in real-time. The proposal of dynamic lexicons instead of static lexicons also solved the problem of recognizing various words in the model by decreasing the number of out o vocabulary words. The increasing text corps also made the resulting language model more accurate. In the future, we will work to increase the speech corpus from 43 hours, enlarge the text corpus from 300 million words, and propose new transformer techniques for end-to-end acoustic modeling without using a separate language model. Using LSTM, BERT, RoBERTa, and GP3 techniques for the language model are other future directions of this research. As well as applying DNN techniques for feature extraction from speech signals is another idea to improve.

REFERENCES

[1] Wang, D., Wang, X. and Lv, S., 2019. An overview of end-to-end automatic speech recognition. Symmetry, 11(8), p.1018.

[2] Wang, Y., Mohamed, A., Le, D., Liu, C., Xiao, A., Mahadeokar, J., Huang, H., Tjandra, A., Zhang, X., Zhang, F. and Fuegen, C., 2020, May. Transformer-based acoustic modeling for hybrid speech recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6874-6878). IEEE.

[3] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxceleB2: Deep speaker recognition," in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2018, vol. 2018-September. doi: 10.21437/Interspeech.2018-1929.

[4] Khalil, R.A., Jones, E., Babar, M.I., Jan, T., Zafar, M.H. and Alhussain, T., 2019. Speech emotion recognition using deep learning techniques: A review. IEEE Access, 7, pp.117327-117345.

[5] M. Johnson et al., "A systematic review of speech recognition technology in health care," BMC Medical Informatics and Decision Making, vol. 14, no. 1. 2014. doi: 10.1186/1472-6947-14-94.

[6] Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T. and Lindén, K., 2019. Automatic language identification in texts: A survey. Journal of Artificial Intelligence Research, 65, pp.675-782.

[7] Tursunov, A., Choeh, J.Y. and Kwon, S., 2021. Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. Sensors, 21(17), p.5892.

[8] Deshmukh, A.M., 2020. Comparison of hidden markov model and recurrent neural network in automatic speech recognition. European Journal of Engineering and Technology Research, 5(8), pp.958-965.

[9] D. Amodei et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in 33rd International Conference on Machine Learning, ICML 2016, 2016, vol. 1.

[10]Y. Xie, L. Le, Y. Zhou, and V. V. Raghavan, "Deep Learning for Natural Language Processing," Handbook of Statistics, vol. 38, pp. 317–328, Jan. 2018, doi: 10.1016/BS.HOST.2018.05.001.

[11]S. M. Omer, J. A. Qadir, and Z. K. Abdul, "Uttered Kurdish digit recognition system," Journal of University of Raparin, vol. 6, no. 2, 2019, doi: 10.26750/vol(6).no(2).paper5.

[12]J. A. Qadir, A. K. Al-Talabani, and H. A. Aziz, "Isolated Spoken Word Recognition Using One-Dimensional Convolutional Neural Network," International Journal of Fuzzy Logic and Intelligent Systems, vol. 20, no. 4, 2020, doi: 10.5391/IJFIS.2020.20.4.272.

[13]Z. K. Abdul, "Kurdish Spoken Letter Recognition based on k-NN and SVM Model," Journal of University of Raparin, vol. 7, no. 4, 2020, doi: 10.26750/vol(7).no(4).paper1.

[14]H. Veisi, H. Hosseini, M. Mohammadamini, W. Fathy, and A. Mahmudi, "Jira: a Kurdish Speech Recognition System Designing and Building Speech Corpus and Pronunciation Lexicon," arXiv preprint arXiv:2102.07412, no. Furui 2005, 2021.

[15]A. Hannun et al., "Deep speech: Scaling up end-to-end speech recognition," arxiv.org.

[16]W. Song and J. Cai, "End-to-End Deep Neural Network for Automatic Speech Recognition," CS224N Projects, 2015.

[17]… P. L.-2019 34th I. and undefined 2019, "Speech recognition using deep learning," ieeexplore.ieee.org.

[18]K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," Applied Intelligence, vol. 42, no. 4, 2015, doi: 10.1007/s10489-014-0629-7.

[19]U. A. KIMANUKA and O. BUYUK, "Turkish Speech Recognition Based On Deep Neural Networks," Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, vol. 22, no. Özel, 2018, doi: 10.19113/sdufbed.12798.

[20]H. Veisi and A. Haji Mani, "Persian speech recognition using deep learning," International Journal of Speech Technology, vol. 23, no. 4, 2020, doi: 10.1007/s10772-020-09768-x.

[21]H. A. Alsayadi, A. A. Abdelhamid, I. Hegazy, and Z. T. Fayed, "Arabic speech recognition using end-to-end deep learning," IET Signal Processing, vol. 15, no. 8, 2021, doi: 10.1049/sil2.12057.

[22]L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," Mar. 2010.

[23]A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in ACM International Conference Proceeding Series, 2006, vol. 148. doi: 10.1145/1143844.1143891.

[24]de la Fuente Garcia, S., Ritchie, C.W. and Luz, S., 2020. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. Journal of Alzheimer's Disease, 78(4), pp.1547-1574.

[25] Yu, D. and Deng, L., 2016. Automatic speech recognition (Vol. 1). Berlin: Springer.

[26] Markovnikov, N., Kipyatkova, I. and Lyakso, E., 2018, September. End-to-end speech recognition in Russian. In International Conference on Speech and Computer (pp. 377-386). Springer, Cham.

[27] Cabral, F.S., Fukai, H. and Tamura, S., 2019. Feature extraction methods proposed for speech recognition are effective on road condition monitoring using smartphone inertial sensors. Sensors, 19(16), p.3481.

[28] Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.Y., 2019. Fastspeech: Fast, robust and controllable text to speech. Advances in Neural Information Processing Systems, 32.

[29] F. S. Cabral, H. Fukai, and S. Tamura, "Feature extraction methods proposed for speech recognition are effective on road condition monitoring using smartphone inertial sensors," Sensors (Switzerland), vol. 19, no. 16, 2019, doi: 10.3390/s19163481.

[30] H. Veisi, M. MohammadAmini, and H. Hosseini, "Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus," Digit. Scholarsh. Humanit., 2019, doi: 10.1093/llc/fqy074.

P- ISSN 1991-8941 E-ISSN 2706-6703 *Journal of University of Anbar for Pure Science (JUAPS)* Open Access
2022,16, (2) :108-118

التعرف التلقائي على الكلام الكردي المركزي باستخدام التعلم العميق

عبد الهادي عباس عبدالله [1] ، هادي فيسي[2]

[1] قسم علوم الحاسوب - كلية العلوم - جامعة سوران سوران أربيل كردستان العراق
aaa160h@cs.soran.edu.iq
[2]جامعة طهران (زائر في جامعة سوران) ، كلية العلوم والتقنيات الجديدة
h.veisi@ut.ac.ir،
Hadi.veisi@Visitor.soran.edu.iq

الخلاصة :

يتم استخدام التعرف التلقائي على الكلام (ASR) باعتباره مجالًا مثيرًا للاهتمام لمعالجة الكلام ، في الوقت الحاضر في التطبيقات الحقيقية التي يتم تنفيذها باستخدام تقنيات مختلفة من بينها الشبكة العصبية الاصطناعية هي الأكثر شيوعًا. تعد زيادة الأداء وجعل هذه الأنظمة قوية في مواجهة الضوضاء من بين التحديات الحالية. تتناول هذه الورقة تطوير نظام anASR للغة الكردية المركزية (CKB) باستخدام نقل التعلم من الشبكات العصبية العميقة .(DNN) مزيج من معاملات Cepstral ذات التردد الميل (MFCCs) لاستخراج ميزات إشارات الكلام ، يتم استخدام طبقة إخراج الذاكرة طويلة المدى (LSTM) مع طبقة إخراج التصنيف الزمني (CTC) لإنشاء نموذج صوتي (AM) على خطاب AsoSoft CKB مجموعة البيانات. أيضًا ، استخدمنا نموذج اللغة N-gram في مجموعة البيانات النصية الكبيرة التي تم جمعها والتي تتضمن حوالي 300 مليون رمز مميز. يتم استخدام مجموعة النصأيضًا لاستخراج نموذج معجم ديناميكي يحتوي على أكثر من 2.5 مليون كلمة .CKB تظهر النتائج التي تم الحصول عليها أن استخدام DNN يحسن النتائج مقارنة بوحدات الإحصاء الكلاسيكية. الطريقة المقترحة تحقق معدل خطأ في الكلمات بنسبة 0.22٪ من خلال الجمع بين تعلم التحويل وتكييف نموذج اللغة. هذه النتيجة أفضل من أفضل نتيجة تم الإبلاغ عنها لـ CKB.