Journal of Techniques, ISSN: 2708-8383, Vol. 5, No. 1, 2023, Pages 176-183
DOI: https://doi.org/10.51173/jt.v5i1.749

JOURNAL OF TECHNIQUES

Journal homepage: http://journal.mtu.edu.iq

RESEARCH ARTICLE - MANAGEMENT

Arabic Speech Recognition Based on Encoder-Decoder Architecture of Transformer

Mohanad Sameer [1*], Ahmed Talib [1], Alla Hussein [2], Husniza Husni [3]

[1] Technical College of Management - Baghdad, Middle Technical University, Baghdad, Iraq

[2] Technical Institute / Kut, Middle Technical University, Baghdad, Iraq

[3] School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

[*] Corresponding author E-mail: dac0014@mtu.edu.iq

Article Info.

Article history:

Received
20 July 2022

Accepted
01 September 2022

Publishing
03 April 2023

Abstract

Recognizing and transcribing human speech has become an increasingly important task. Recently, researchers have been more interested in automatic speech recognition (ASR) using End to End models. Previous choices for the Arabic ASR architecture have been time-delay neural networks, recurrent neural networks (RNN), and long short-term memory (LSTM). Preview end-to-end approaches have suffered from slow training and inference speed because of the limitations of training parallelization, and they require a large amount of data to achieve acceptable results in recognizing Arabic speech. This research presents an Arabic speech recognition based on a transformer encoder-decoder architecture with self-attention to transcribe Arabic audio speech segments into text, which can be trained faster with more efficiency. The proposed model exceeds the performance of previous end-to-end approaches when utilizing the Common Voice dataset from Mozilla. In this research, we introduced a speech-transformer model that was trained over 110 epochs using only 112 hours of speech. Although Arabic is considered one of the languages that are difficult to interpret by speech recognition systems, we achieved the best word error rate (WER) of 3.2 compared to other systems whose training requires a very large amount of data. The proposed system was evaluated on the common voice 8.0 dataset without using the language model.

Publisher : Middle Technical University

Keywords: Sequence to Sequence ASR; Arabic ASR; Transformer-Speech Recognition; Arabic Speech to Text.

1. Introduction

Automatic speech recognition for Arabic is a very challenging task because of the complex morphological nature of the language and the absence of short vowels in written text, which leads to several potential vowelizations for each grapheme, which are often conflicting. The primary goal of ASR is the machine's ability to interpret human speech and translate it to text or readable form with the ability to implement an operation based on instructions set by a human. Over the past decades, ASR technologies have played multiple roles in many fields. For example, education, personal computers, robotics, mobile phones, armies, health, security systems...etc. Since deep learning for ASR systems has been available [1], multiple neural network model architectures for ASR systems have been proposed in [2-6], The architectures mentioned above include recurrent neural networks (RNN), particularly LSTM [7]. neural networks are frequently employed whether, in the traditional system [3, 8], sequence-to-sequence-based systems [9, 10]), or end-to-end systems based on tranducers [11]. The RNNs on the other hand has some famous restrictions:

1) RNNs cannot deal with long-term temporal dependencies well Because of the problem of exploding and vanishing gradients identified in [8].
2) The recurrence nature of RNNs makes the processing of speech signals in parallel difficult.

To solve these problems, different architectures have been presented to replace RNNs, including TDNN [5], feed-forward sequential memory networks FSMN [6], and CNN [4, 9], since just limited improvement has been reached in the results. It is worth mentioning that there is little research on the recognition of spoken speech in the Arabic language compared to other languages [10] noting that the Arabic language is one of the 5 most important languages [11] and its speakers are more than 400 million people, and it is the language of the Holy Qur'an, which believed by more than 1.8 billion people around the world in 2015 and the number is expected to be 3 billion in 2060.[12] The Arabic language is one of the oldest languages in the world. It is a language with great diversity. The Arabic language can be divided into three types:

1) Classical Arabic: It is the language of the Islamic religion (the language of the Hadith and Qur'an, and the language of ancient Arabic poetry).
2) Modern Standard Arabic (MSA): Standard Arabic is utilized in communication, official correspondence newspapers, news, and modern books.
3) Dialectal Arabic: Dialectal Arabic is used for everyday informal communication and has many forms(accents).

Nomenclature & Symbols			
CNN	Convolution Neural Network	LSTM	Long Short-Term Memory
MSA	Modern Standard Arabic	GMM	Gaussian Mixture Model
RNN	Recurrent Neural Network	BDRNN	Bidirectional Neural Network
URL	Uniform Resource Locator	DNN	Deep Neural Network
WER	Word Error Rate	MFCC	Mel-Frequency Cepstral Coefficients
TDNN	Time Delay Neural Networks	FSMN	Feed-Forward Sequential Memory Networks
d_{model}	Dimensional Output	HMM	Hidden Markov Model

This paper aims to introduce an ASR system for recognizing Arabic speech using a Transformer based on the attention mechanism technique to recognize modern standard Arabic speech and show how well the proposed model performs in a truly low-resource language as modern standard Arabic and tries to achieve the best possible WER in Speech recognition by using the available data.

Our contributions are as follows. First, we show that depth is an important factor to acquire competitive end-to-end ASR models with the Transformer especially for recognizing Arabic speech. Second, obtain the state-of-the-art result among end-to-end ASR models for the Arabic Common voice 8.0. This result is achieved using a total of 5 encoder layers,3 decoders, and 460 feed-forward layers. Note that the best result was previously achieved at 40.6% of WER on the common voice dataset [13].

2. Related Works

Speech recognition enables machines to recognize and respond to human speech. ASR is a machine's ability to recognize (interpret) spoken language and convert it into a textual form, as well as take actions based on human-defined instructions [14]. In [15] The authors reported their experiences in building the Arabic version of CMU's Sphinx4 ASR system, where they presented a system to recognize the speech of Arabic digits. In [16]the authors proposed a dataset of Algerian Arabic speech, consisting of statements spoken by 300 Algerian speakers with different accents, picked from 11 areas of Algeria.

In [17] the authors introduced a technique of Arabic ASR based on Sphinx-Train and also introduced a corpus of 11.5 hours of labeled speech data for system training, their voice recognition system performance was evaluated using three Arabic speakers only. The results achieved around a 97% recognition rate for Arabic single words. [18] proposed a system for phoneme recognition as an ASR system. Many techniques are used in this paper Firstly, in the stage of processing, the algorithm of Gaussians' Low-Pass-Filtering was used with an artificial neural network to improve the result. In the recognition of the phoneme stage, signal catching, sampling, quantization, and setting energy are done, then, a neural network is utilized to improve the achieved result. When using the Gaussians' Low-Pass-Filtering on voice signals, the results have an enhanced impact due to noise reduction, after that, the neural network will be used in the training phase to recognize the speech signal. The authors in [19] presented an architecture for Arabic ASR, which consists of four stages:

1) Pre-processing stage.
2) Feature extraction stage.
3) Pronunciation dictionary, language model, and Acoustic model (decoding).
4) Stage of Post-processing result in which the best hypotheses are produced.

The working mechanism of these stages is as follows:

Pre-processing stage inputs are utterance speech signals, then the output is processed voice signals used as input to the feature extraction stage, where the features are represented by vectors as output. After that, the vectors are used as input in the decoding stage, and it works along with a pronunciation dictionary. Finally, the outputs from the pronunciation dictionary step are fed into the post-processing step. In the hypothesis-search component, acoustic and language models are combined, scores are given to the features vector sequence and the predicted word sequence, then the word sequence with the best score is produced. In [10], the authors proposed three approaches to improve the ASR system. The first approach in pronunciation modeling is the employment of a decision tree with pronunciation variant generation. Then, a hybrid method is presented to adapt the native acoustic model with another native acoustic model. Lastly, processed text was used to improve the language model. In the result, the word error rate is reduced by 1%, 1.2%, and 1.9% for the pronunciation model, the acoustic modeling, and the language model, respectively.

The Hidden Markov Model and Gaussian Mixture Model have been widely used for a long time in large-vocabulary continuous speech recognition (LVCSR).[20] introduced the first HMM–DNN hybrid technique, in which the GMM model was replaced by a deep neural network model. LVCSR performance was significantly improved compared to the previous HMM-GMM systems. After that, for acoustic modeling, several DNN architectures were investigated, such as Recurrent Neural Networks (RNN), Bidirectional RNN(BDRNN), and deep conditional random fields, all of which demonstrated a significant increase in performance [21, 22]. In [5], the researchers introduced a time-delay neural network and obtained higher performance in learning and broader temporal dependencies than DNN- RNN-based architectures. In [11], the authors presented the recognition of isolated Arabic spoken words that were used in two ASR applications(Television voice command recognition and spoken digit recognition). System components include (signal acquisition, feature extraction using MFCCs, corpus construction, model training with RNN and gated RNN, and classification). In [13], the authors attempted to build a sequence-to-sequence ASR system for the Tunisian dialect and modern standard Arabic(Common Voice dataset) by using deep learning algorithms based on DeepSpeech2 architecture, they built a Tunisian speech dialect corpus named "TunSpeech" which consisted of 11 hours only and achieved 40.6% WER on the common voice that they trained their system on, Whereas, in this research, we achieved an accuracy of performance that reached 3.2 of WER on the Common Voice dataset using our speech-transformer model. In [23], the authors proposed an ASR system by utilizing a deep neural network-based hybrid and a Transformer-based End to End model to build the first ASR system and releasing the first speech corpus for the Egyptian Arabic–English language pair in which hypotheses of the two systems are merged at the sentence and word level. These techniques

result in a total WER relative enhancement of 4.7% over the original WER score of 32.1%. In the instance of intra-sentential CS sentences, they obtain a word error rate significant enhancement of 4.8 %. and the best system performance achieved a word error rate of 30.6% on the ArzEn test set.

Finally, we would like to clarify here that the proposed system using the (Transformer) is the first of its kind that has been trained and tested on a common voice dataset. In Fig. 1, we illustrate the historical development of speech recognition systems from 1950 to the recent and important developments that have occurred over the years.

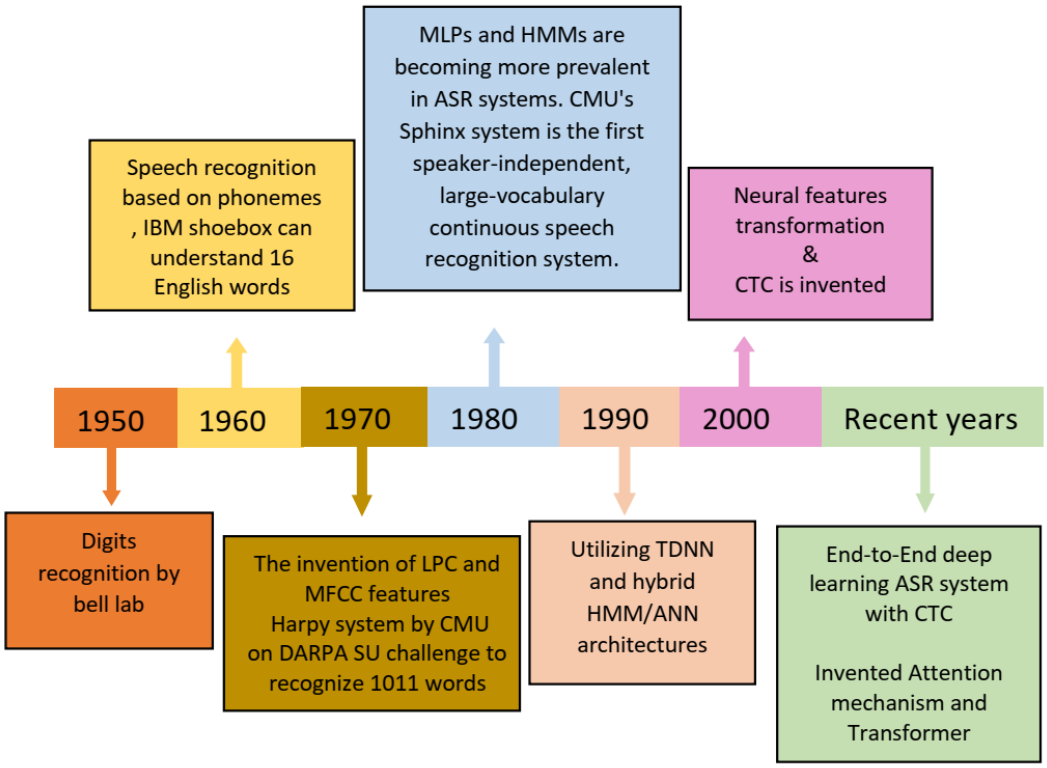

Fig. 1. Historical Development of Speech Recognition Systems

3. Proposed Model

In this section, we will first explain the pre-processing operations on the dataset, followed by the Methodology that describes the Acoustic Input Layer of the system, and then explains in detail the Architecture of the Transformer Model. Finally, the most important components of the transformer-speech recognition system are stated.

3.1. Pre-Processing

The sources of train, dev, and test are all data that have been reviewed and deemed of good quality. We cleaned the data by using only the allowed set of characters Arabic language. All audio clips are in an original dataset in mp3 format, and then we converted the format of audio from mp3 to Wav with a 16000-sample rate, a mono audio channel, and a depth of 16 bits.
Regular expression rules and processes were used to correct the orthographic problems in the dataset. The following are some steps used to reduce ambiguity in spelling and pronunciation:

- Remove all URLs, E-mail addresses, paths, and special characters (&, @, $,;,:, ., (), ",….) pronunciation, and non-Arabic texts.
- All diacritics representing consonant stressing or short vowels are stripped.
- All numbers are normalized and they are converted into literal words.
- Remove any double space between the words
- The stretched words are reduced to their original form. For example, الرجاااال is replaced by الرجال "men"
- If a ة ta marobuTa is connected to the next word, a space is added after each word. For example,نهايةالقرن is replaced by نهاية القرن "century end".
- The time is expressed in a literal manner, as shown for instance 15:30 is replaced by الثالثة و ثلاثون دقيقة
- Replace some abbreviations with their corresponding meaning (As shown in Table 1).

3.2. Architecture of Transformer Model

We apply four convolutional layers to downsample acoustic features (through convolution strides) and process local relations and calculate the sum of the token and position embeddings when processing past target tokens for the decoder.
The proposed speech recognition-Transformer model aims to transform the speech features sequence into the corresponding characters sequence. It can be represented by a two-dimensional spectrogram with time and frequency axes.

Table 1. Abbreviations of Arabic words

Abbreviation	Arabic words	English term
%	في المائة	percent
ت غ	توقيت غرينتش	GMT
ت	تاريخ	Date
هـ	هجري	Islamic Calendar
د. ك	دولار كندي	Canadian dollar
س	ساعة	Hour
د	دقيقة	Minute
ث	ثانية	Second

3.2.1. Encoder-Decoder Block

The model consists of two important parts. The encoder is the first, consuming the source sequence and producing a high-level representation; the second is the decoder, which generates the target sequence.

Encoder and Decoder are both neural networks that consist of neural parts that can learn the relationship between the input and output time steps. The decoder also includes the process to condition particular encoder representation components. Instead of recurrence (RNN), multi-head attention or its attention mechanism is the heart of the Transformer model.

Input audio spectrograms are used by our model to predict a character sequence. The target character sequence is presented to the decoder during training. The decoder predicts the following token during inference by using its prior predictions [24, 25], see Fig. 2.

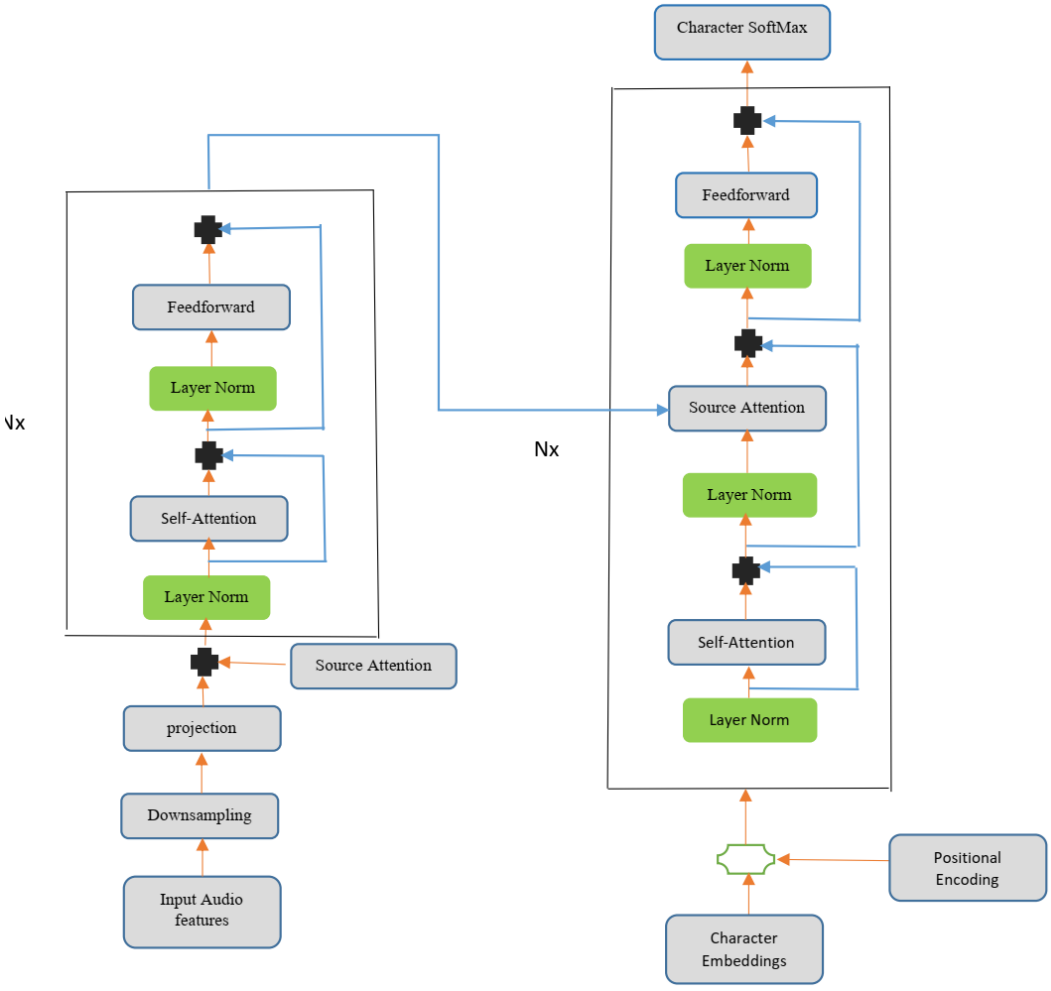

Fig. 2. Speech-Transformer model architecture

3.2.2. Attention

Attention is the process of employing a content-based information extractor from a set of Q queries, K keys, and V values. Similarities between the Q and the K, and in turn return the weighted sum of the values using Scaled Dot-product-attention are the basis for the retrieval function

[26]. Scaled Dot Product Attention is an efficient method for self-attention proposed in [25] As illustrated in Fig. 3, Let Q ∈ R×d_q are queries, K ∈ Rtq×dq are keys and V ∈ Rtq×dq are valued, where t represents the element numbers in various inputs and d∗ is the corresponding element dimension. Normally, $t_k = t_v$, $d_q = d_k$. The outputs of self-attention are calculated as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

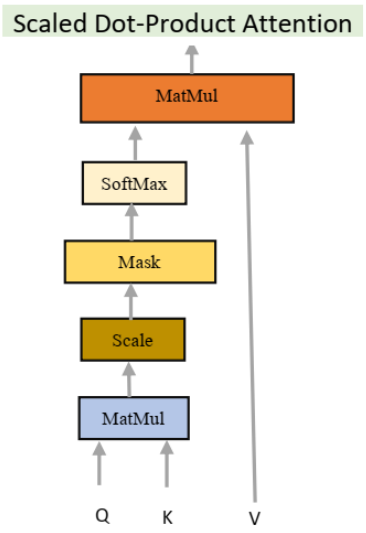

Fig. 3. Scaled Dot-Product Attention

Where:
MatMul is a function that returns the matrix product of two arrays, SoftMax is an activation function, mask refers to masking it is utilized to ensure the predictions for position j can depend only on the known outputs at positions less than j, Scale is meaning scale down where the scaler $1/\sqrt{d_k}$ is used to prevent SoftMax function into regions that have very small gradients.

3.2.3. Multi-Head Attention

The transformer is consisting of multiple dot Attention layers [26], in [25] dot-product attention has recently been improved, which scales the queries before and then introduces a sub-space projection for K, Q, and V into n parallel heads. Since n-attention operations are implemented with corresponding heads for each operation. The results are the chain from the attention output of each head. Conspicuously, different from recurrent connections which use a single state with a gating structure for Data transmissions, or convolution connections that linearly combine local states constrained by a kernel size, self-attention is an aggregation of the information in all time steps without intermediate transformation. The Scaled Dot-Product Attention is computed individually, and their outputs are sequenced and fed into other linear projections to produce the final dimensional outputs d_{model} , see Fig. 4.

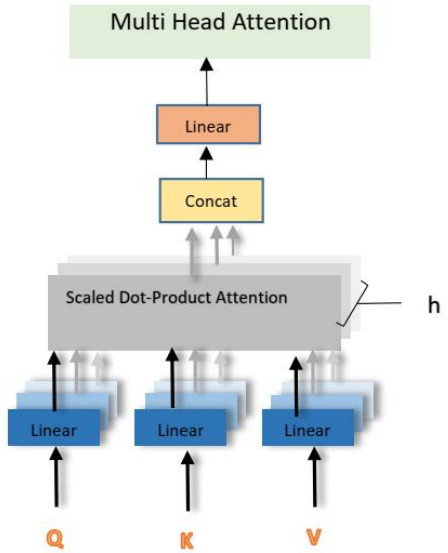

Fig. 4. Multi-Head Attention architecture

3.3. Speech Model and Layers Description

The Speech recognition using Transformer is like the previous sequence-to-sequence models, the encoder transforms a speech features sequence ($x_1, \ldots, x_T$) to a hidden representation $h = (h_1, \ldots, h_L)$, then the decoder produces an output sequence ($y_1,\ldots, y_s$). the decoder uses the prior emitted characters at each step as additional inputs when emitting the next character [25].

As illustrated in Fig. 1. In Transformers, the encoder-decoder layers are made up of Self Attentional sub-layers that are connected with feed-forward neural layers. To accommodate extended speech statements, we adopt the reshaping technique described in [6] by gathering consecutive frames into a single step, then, we integrate the acoustic features with sinusoidal positional encoding [25, 27, 28]. The input features from the Transformer encoder are passed to a self-attention layer, then to a feed-forward neural network with one hidden layer that uses the activation function of ReLU. Before these sub-modules, we include residual connections that create shortcuts between the higher layer and the lower-level representation. The normalization layers positioned after each residual connection significantly reduce the amplitude of the neuron values caused by the residual layer's presence [29]. Modern translation systems use the Transformer decoder as the original transformer decoder from [25]. To keep the model's auto-regressive nature, the decoder's Self Attention layer should be masked such that each state only has access to the previous states. This is a significant distinction between the decoder and the encoder. Moreover, Between the self-attention and the feed-forward layers, there is an additional attention layer that uses the target hidden layers as queries and the encoder outputs as keys and values. Comparing this particular Transformer design to previous proposed RNNs and CNNs networks, there are multiple advantages, such that each layer and sub-module can be efficiently parallelized over the mini-batch and time dimensions of input data.in addition to the combination of residual and layer normalization are the keys to allowing more depth configurations to be trainable. This is the primary factor behind the result performance improvement in current research in both machine translation and NLP [30, 31].

4. Experimental Result

Mozilla's Common Voice is an audio collection that intends to make the recognition of human speech accessible to everyone. Common Voice dataset includes English, Arabic, French, and German, in addition to more than ten other languages. The corpora were created outside of a controlled environment and setting. The speech may contain noise in the background, and users may have diverse accents. Common Voice consists of a unique MP3 and corresponding text file. The Arabic Common Voice version 8.0 contained 139 recorded hours in the dataset (before preprocessing). Afterward, we trained the model for 112 hours (after preprocessing the dataset and skipping the transcription that contained vocabulary words and characters). The dataset consists of demographic metadata such as gender, accent, and age. May assist in training speech recognition engines [36].

Textual data is data from texts utilized to create the language model and to improve Deep Speech results. We use one of the largest available textual corpus is a raw text from Arabic daily newspapers collected over a year between 2004 and 2005 by Ahmed Abdelali that contains 250354 lines of statements and 2941404 words.

The Python language was used to program the system based on the libraries of TensorFlow, Keras, NumPy, Glob, Os, Pydub, Wave, Audioop, and Jiwer which were used to obtain the Word Error Rate of the predictions, pyyaml h5py that required for saving models and checkpoints due to training in HDF5 format. The Google Colab Pro was used to train the model as illustrated in Fig. 5. It offers GPU (NVIDIA-SMI 460.32.03, GPU name is Persistence-M) between 50 and 120 epochs using a custom learning rate schedule, and an early stop is used where the validation loss is not improved for 6 epochs. MSA common voice 8.0 dataset was utilized for train validation, and testing with splitting rates of 90%, 5%, and 5%. To train the system with the best combination and achieve the best result, we have conducted several experiments to train the system by using different variables. The best results were achieved when using frame-length=200, frame-step=80, and a batch size of training is 128, Finally, the duration of the training and testing of the model was about 34 hours.

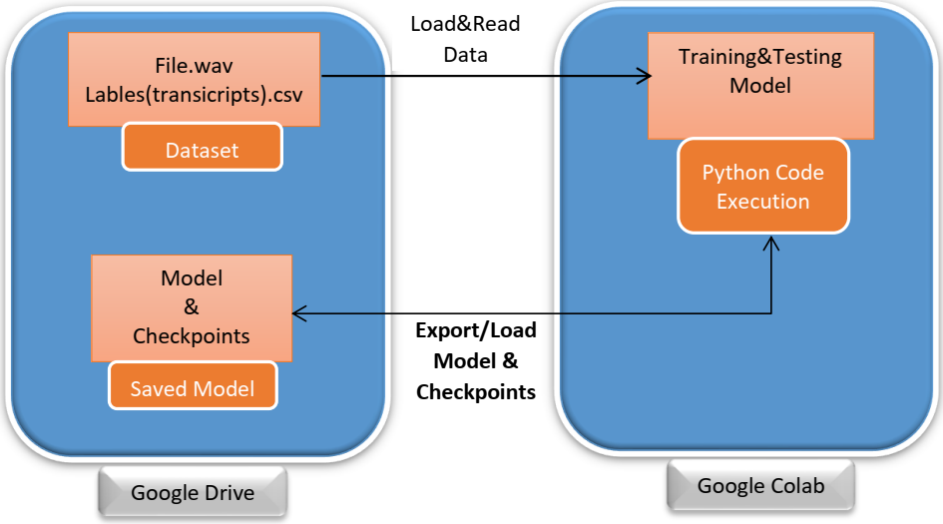

Fig. 5. Speech recognition training execution on Colab

The word error rate, which is a widely used measurement in automatic speech recognition evaluation, is defined as follows:

$$WER = \frac{S+D+I}{N} \quad (7)$$

N: number of words in the source sentence.
S: substituted words
D: Deleted word

I: Inserted word
Note that increasing insertion may cause WER to be greater than 1.

5. Results and Discussion

We have used different hyperparameters to train the Arabic speech recognition model to achieve the best-desired results by changing the number of encoders N_e and the number of decoders N_d as we mention in Tables 2 and 3, ignoring experiments that led to the overfitting or underfitting.

Table 2. Results Summary

Model No.	N_e	N_d	Feedforward layer	WER %
1	5	2	400	3.9
2	5	3	400	3.7
3	6	3	400	5.8
4	5	3	460	3.2
5	6	5	460	6.2
6	8	6	460	6.3

Table 3. Examples of Prediction

Target	Prediction
انا اعرف عنوانه	انا اعرف عنوانه
هل تذكر الليلة التي تقابلنا فيها اول مرة	هل تذكر الليلة التي تقابلنا فيها اول مرة
هو منزل النجوى بخالي الأعصر	وهو منزل النجوى بخالي الأعصر
لا احب ان اكون وحدي	لا احب ان كل وحدي
يقود أبي السيارة إلى عمله	يقود أبي السيارة إلى عمله
أبا حفيص رويدا	أبا حفيس رويدا

The best results were achieved through the use of 5 encoders and 3 decoders, obtaining a word error rate of 3.2. We can compare this result with the latest result previously achieved by Messaoudi et al. (2021)[13], where the word error rate was achieved at 4.0 by utilizing the common voice dataset.

When analyzing the errors made by the proposed model, one issue that stands out is a truncated output. Quite a lot of output texts are much shorter than the source transcripts. The performance of the model will be seriously impacted by the failure to produce long enough output sentences. Automatic speech recognition has a lot of challenges, specifically about spontaneous Arabic speech due to the prevalence of "uh" and "um" disfluencies, as well as other speech irregularities such as stuttering and coughing. Therefore, the system must be capable of dealing with out-of-vocabulary words (OOV). Other issues are the variation of speakers, word speed, accent, pitch, etc.

6. Conclusion

In this paper, we proposed an ASR system to recognize the Arabic language based on a Transformer that depends on Attention-Mechanism with positional encoding to learn position dependencies. The proposed work achieved the best accuracy in recognizing Arabic speech and reduced the heavy training cost compared to the preview recurrence models. The proposed Speech-Transformer system achieved the state-of-the-art performance 3.2 of WER on Common Voice when evaluated on the Common Voice dataset using five encoders and three decoders. We trained the ASR model on the Modern Standard Arabic language without using a language model and utilizing only 112 hours for training and evaluating the system. In future work, the system can be trained on more data to get better results, and it can also be trained on the Iraqi dialect or mixed MSA-Iraqi dialect to achieve the optimum benefit from speech recognition applications in Iraqi communities, organizations, institutions, and companies.

Acknowledgment

This work is extracted from a thesis; it was submitted as part of the requirements for obtaining a master's degree in managerial informatics techniques. I took this opportunity to thank everyone who helped and supported me to complete this work so that it would reflect the scientific standing of my esteemed university.

References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal processing magazine, vol. 29, pp. 82-97, 2012.
[2] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in Twelfth annual conference of the international speech communication association, 2011.
[3] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.
[4] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," IEEE/ACM Transactions on audio, speech, and language processing, vol. 22, pp. 1533-1545, 2014.

[5] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in Sixteenth annual conference of the international speech communication association, 2015.

[6] S. Zhang, H. Jiang, S. Wei, and L. Dai, "Feedforward sequential memory neural networks without recurrent feedback," arXiv preprint arXiv:1510.02693, 2015.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, pp. 1735-1780, 1997.

[8] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE transactions on neural networks, vol. 5, pp. 157-166, 1994.

[9] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," arXiv preprint arXiv:1609.03193, 2016.

[10] B. H. Ahmed and A. S. Ghabayen, "Arabic automatic speech recognition enhancement," in 2017 Palestinian International Conference on Information and Communication Technology (PICICT), 2017, pp. 98-102.

[11] 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.

[12] A. A. Abdelhamid, H. A. Alsayadi, I. Hegazy, and Z. T. Fayed, "End-to-end arabic speech recognition: A review," in Proceedings of the 19th Conference of Language Engineering (ESOLEC'19), Alexandria, Egypt, 2020, pp. 26-30.

[13] A. Messaoudi, H. Haddad, C. Fourati, M. B. Hmida, A. B. E. Mabrouk, and M. Graiet, "Tunisian Dialectal End-to-end Speech Recognition based on DeepSpeech," Procedia Computer Science, vol. 189, pp. 183-190, 2021.

[14] H. H. Nasereddin and A. A. R. Omari, "Classification techniques for automatic speech recognition (ASR) algorithms used with real time speech translation," in 2017 Computing Conference, 2017, pp. 200-207.

[15] H. Satori, M. Harti, and N. Chenfour, "Introduction to Arabic speech recognition using CMUSphinx system," arXiv preprint arXiv:0704.2083, 2007.

[16] S. A. Selouani and M. Boudraa, "Algerian Arabic speech database (ALGASD): corpus design and automatic speech recognition application," Arabian Journal for Science and Engineering, vol. 35, pp. 157-166, 2010.

[17] M. Belgacem, A. Maatallaoui, and M. Zrigui, "Arabic language learning assistance based on automatic speech recognition system," in Proceedings of the International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE), 2011, p. 1.

[18] K. Khatatneh, "A novel Arabic Speech Recognition method using neural networks and Gaussian Filtering," International Journal of Electrical, Electronics & Computer Systems, vol. 19, 2014.

[19] D. Yu and L. Deng, Automatic speech recognition vol. 1: Springer, 2016.

[20] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Transactions on audio, speech, and language processing, vol. 20, pp. 30-42, 2011.

[21] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in 2013 IEEE international conference on acoustics, speech, and signal processing, 2013, pp. 6645-6649.

[22] Y. Hifny, "Unified acoustic modeling using deep conditional random fields," Transactions on Machine Learning and Artificial Intelligence, vol. 3, p. 65, 2015.

[23] I. Hamed, P. Denisov, C.-Y. Li, M. Elmahdy, S. Abdennadher, and N. T. Vu, "Investigations on speech recognition systems for low-resource dialectal Arabic–English code-switching speech," Computer Speech & Language, vol. 72, p. 101278, 2022/03/01/ 2022.

[24] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5884-5888.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[26] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," arXiv preprint arXiv:1508.04025, 2015.

[27] M. Sperber, J. Niehues, G. Neubig, S. Stüker, and A. Waibel, "Self-attentional acoustic models," arXiv preprint arXiv:1803.09519, 2018.

[28] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in International conference on machine learning, 2017, pp. 1243-1252.

[29] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv preprint arXiv:1607.06450, 2016.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[31] N.-Q. Pham, J. Niehues, T.-L. Ha, E. Cho, M. Sperber, and A. Waibel, "The karlsruhe institute of technology systems for the news translation task in wmt 2017," in Proceedings of the Second Conference on Machine Translation, 2017, pp. 366-373.

[32] A. Bapna, M. X. Chen, O. Firat, Y. Cao, and Y. Wu, "Training deeper neural machine translation models with transparent attention," arXiv preprint arXiv:1808.07561, 2018.

[33] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," Advances in neural information processing systems, vol. 29, 2016.

[34] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in European conference on computer vision, 2016, pp. 646-661.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," The journal of machine learning research, vol. 15, pp. 1929-1958, 2014.

[36] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, et al., "Common voice: A massively-multilingual speech corpus," arXiv preprint arXiv:1912.06670, 2019.