Thyroid Ultrasound Report Generation Based On Urdu Speech Recognition

Naeem Haider*, Wenzhong Yang*[1]

[*] School of Information Science and Engineering, Xinjiang University, China

[1] For correspondence: Wenzhong Yang (xwz_xy@163.com)

DOI: 10.29322/IJSRP.12.07.2022.p12703
http://dx.doi.org/10.29322/IJSRP.12.07.2022.p12703

Paper Received Date: 9th June 2022
Paper Acceptance Date: 28th June 2022
Paper Publication Date: 6th July 2022

Abstract- Medical reports generations especially in a third world countries can take unnecessary long hours. Efficient speech recognition technology trained under medical lexicon can greatly help not only the physicians but also the patients by reducing a lot of time. In order to build a robust speech recognition model that can not only differentiate and understand medical terminologies along with other generic terms and utterances but also grasps the nuances and structures of very different language. We could not make library of every medical terminology from every kind of medical field, so we chose to limit this research to only thyroid report generation. Secondly, we tried to get the optimal speech recognition results by using our local language Urdu. In order to tackle both the challenges, we used Conformer model along with other acoustic models to identify the best performing structure. CNN and deep CNN model WER (word error rate) was 16.9% and 14.2% respectively, indicating deep CNN performs better. Similarly, after the addition of maxout function in both CNN and deep CNN, the results were better i.e., 16.5% and 13.8% respectively. Furthermore, I integrated CTC with both CNN and deep CNN. WER of CTC-DCNN (maxout) was 12.2% which was better than 15.3% of CTC-CNN (maxout). We used all these models so that we can compare them to the Conformer model which is the most recent model and certainly in this research WER with the use of Conformer was 8.8%. Further adaptations in Conformer can surely improve its accuracy even more.

***Keywords*:** Urdu Speech Recognition, Conformer, Convolutional Neural Network, Deep Convolutional Neural Network, Thyroid Ultrasound Report

I. INTRODUCTION

Speech recognition should not just be limited to major giants in the form of SIRI, UPS, Amazon, IBM or Alexa, it should be available to every field. Test report generation can consume a lot of time and when we consider third world countries it can become cumbersome. Even if we just look at its surface, the implementation scale is humongous. My idea is to give some help to the physicians who make reports on other renowned software which do not recognize Urdu language. This makes their work tiring or they have to employ some other person to do this task. Medical reports generation, from Urdu speech recognition, dictated by physicians will have a huge impact on the people's lives. In this research, I have mainly focused on thyroid report generation so we have trained our model according to that.

Thyroid disorders are increasing in Pakistan. The prevalence was around 10% at the World Thyroid Day 2021. This is not the actual number as most of the cases remain undiagnosed because of poor healthcare facilities and less awareness [1]. Timely diagnosis and detection can help in the cure of thyroid related diseases and cancer as well. In Pakistan, a normal thyroid test ranges between 20-50 USD and it takes 3-14 days to generate results. Conformer based Urdu speech recognition will have huge scale implementations in a country of 229 million people. Speech recognition and its underlying processes and layers have to be discussed to understand how our model works.

TRADITIONAL APPROACHES TO THE SPEECH RECOGNITION SYSTEM

ASR (Automatic speech recognition), which recognizes words in spoken language and transforms them into text, has a wide range of applications which includes command and control, transcribing speech which is recorded, dictation, conversations with interactive speaking and looking for audio documents. We have explained some of the previous models that have been used for speech recognition system.

URDU SPEECH RECOGNITION

Since 2008, Speech recognition primarily based on the Urdu language has started to apply very commonly [2]. In the past studies [3,4], the acoustic model primarily counts on neural networks and has been broadly used in the Urdu language and has performed extremely well. In the near past, HFNs (Hybrid Frame Neural Networks) clearly have made remarkable progress [5], with previously the most successful system achieving a WER score of 6.99 percent. However, research in the Urdu speech recognition field is very much new to the world. In the very beginning, Azam Beg [3] used a neural network for the acoustic model of the Urdu speech recognition system, resulting, significant improvement in the performance of the Urdu speech recognition and its commercialization. Researchers like Hazrat Ali and Nasir Ahmad [6] have conducted extensive studies on several areas of Urdu speech recognition. Javed Ashraf and Naveed Iqbal [7], for example, created Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System [8]. At the experimental stage, they employed an extended vocabulary continuous speech recognition system based largely on HMM that is an advantageous attempt on language rules and acoustic properties. They employ triphones because of the primary sub-word speech unit and a statistical language model, undertake experiments meanwhile assembling HMMs. They have experience building and optimizing corpus in addition to creating models and programming [9,10]. The LVCSR Urdu speech recognition system has a recognition rate of 85.16 percent. Because of the late evolution of this technology of the Urdu speech recognition, the quantity of research papers on Urdu language recognition is comparable to that of other regularly used languages. Pakistan’s Information and Communication Research Institute conducted research on Urdu speech recognition for the first time. Based on the research results, they proposed a multi-class N-gram model suitable for Urdu vocabulary features and successfully established the Urdu speech recognition system [11]. Therefore, through more than 15 years of development, the scope of research on Urdu language recognition has continued to expand and has begun to enter a deeper research stage.

Urdu Grammar, Word, and Morpheme have already been discussed in reference [12]. The DNN model is used to stimulate the acoustic unit’s posterior probability directly. When combined with HMM, the latter fully exploits the advantages of the generative and discriminative models and has a noticeable recognition impact when dealing with today's huge data. It is superior to the former, but its training parameters are huge, and memory requirements are strict. How to resolve these problems, lessen education parameters, and in addition, enhance speech recognition rate is especially prominent. CNN (convolutional neural network), a kind of DNN (deep neural network), has gained growing interest because of the reduction of reliance on statistics via specific model architectures [13]. However, the DNN has strict training parameters; it still cannot achieve the intended impact, despite improving the WER (word error rate). Speeding research into Urdu speech recognition is beneficial to supporting the enhancement of traditional Urdu knowledge and has significant implications for education, transportation, and communication in Urdu. In recent years, there has been a lot of improvements in the ASR systems that were based on neural networks. The de-facto choice for automatic speech recognition have been RNNs (Recurrent Neural Networks) [4,5] as they have the better tendency to model the temporal tendencies [14].

Recently, the Transformer has enjoyed worldwide acceptability for sequencing the models using self-attention. The quality of this architecture lies in its ability to catch the long distanced interactions and training efficiency [15]. On the other hand, ASR works successfully with convolutions. This process uses the capturing of the local context in order through the local receptive field [16,17]. But the above discussed two models have their limitations. Transformers can work better with the extensive long range context but they have limited working when it comes to the extraction of fine-grained local features pattern. But when it comes to CNNs (conventional neural networks), usually local information is exploited and used as de-facto block. You will learn common position-based kernels via a local window, maintaining translation equivariance, and being able to capture features such as edges and shapes. The usage of local connectivity has the limitation when it comes to capturing the global information. It has to use multiple extra parameters/layers to get that global information. To mitigate the problem, squeeze and extinction module is being adopted by the Context Net [18]. It works on each residual block and captures the longer context [19]. But this is still limited as it only captures the global context because it only works on the application of the global average on the complete sequence. As shown in Fig. 1 Conformer consists of two macaron-like feed-forward layers with half-stage residual connections enclosing the multi-headed self-awareness and folding modules. This is followed by a post-layer norm. In the recent research, they have used the combined versions of the self-attention and convolutions and received the better results rather than running them individually [20]. When used together, their work pattern transfers to the way of combining the local and global interactions of position and content features. Simultaneously, some articles for example [12,21] have increased self-attention that maintains equivariance by positioning relatively based on the information. There is another way proposed by Wu *et al.*, [10]. It splits the multi branched architecture’s input into two parts; one is convolutions and secondly self-attention. Their output is again obtained in a single branch; concatenated. The main target of this work was applications used in mobiles and it also showed the improvements in the tasks of machine translation. Combination of transformers and convolution neural networks efficiently model both audio sequence dependencies (local and global).

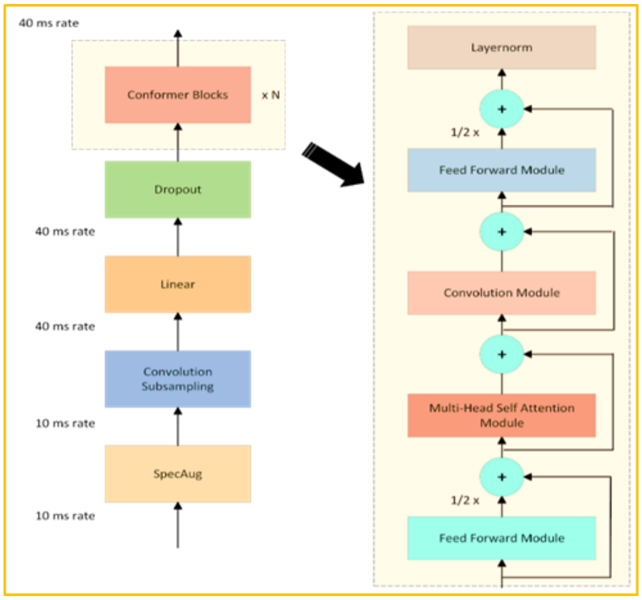

Fig. 1. Conformer encoder model architecture

In the light of all the above research, the Conformer was designed for improved speech recognition. Conformer is a combination of CNN and transformer models. This model has outperformed the previous models based on Transformer and CNN significantly by giving the best till date accuracies. As it was performed on the LibriSpeech, an extensively used benchmark, Conformer achieved WER of 2.1%/4.3% and 1.9%/3.9% without and with the usage of extrinsic language mode respectively on test other. When it was run with a tiny model having 10M parameters, a competitive result of 2.7%/6.3% was observed [22].

I. DESIGN AND IMPLEMENTATION

CONFORMER DESIGN

CONFORMER ENCODER

This encoder works differently from the previous models, as the first step after input is the processing through a layer of convolution subsampling, further processed with multiple conformer blocks. This working is explained in Fig. 1. The thing which distinguishes our model from previous works is the use of Conformer blocks instead of Transformer blocks [15,23]. This block is based on 4 modules working altogether i.e. a convolution module, a self-attention module, a second feed-forward module and a feed-forward module.

MULTI-HEADED SELF-ATTENTION MODULE

MHSA (Multi-headed self-attention) is based on a vital technique from Transformer-XL [24]. This works on such an encoding scheme that considers relative sinusoidal positioning. Relative position encoding allows the self-awareness module to better generalize to different input lengths, and the modified encoder is strong to utterance length variance. They used pre-normalized residual units [25] with dropout, which supports training and regularization of deeper models. Fig. 2 below explains the multi-headed self-awareness block.

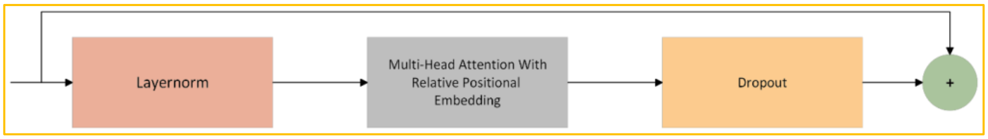

Fig. 2. Multi-Headed self-attention module.

CONVOLUTION MODULE

S. H. Zhanghao *et al.*, [10] inspired to start the convolution module with a gating mechanism which is a GLU(gated linear unit) and a pointwise convolution which follows a single 1-D depth wise convolution layer. In order to aid deep models training, Batchnorm is moved after the convolution. The Convolution block is illustrated in Fig. 3.

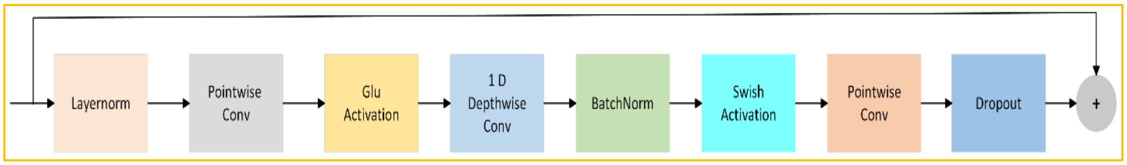

Fig. 3. Convolution module.

FEED FORWARD MODULE (FFN)

According to Vaswani *et al* [26], the architecture of the Transformer is arranged in such a way that MHSA layer is followed by feed forward module. Nonlinear activation is sandwiched between two linear transformations, make this Transformer architecture. After the normalization, addition of a residual connection over the feed-forward layers is structured. Transformer ASR models also adopted this structure [15,27]. Pre-norm residual units [25] are followed and layer normalization is applied within the residual. This is done before the first linear layer on the input. Swish activation [28] and dropout is also applied, which is a great help for the network regularization. Fig. 4. shows the FFN module.

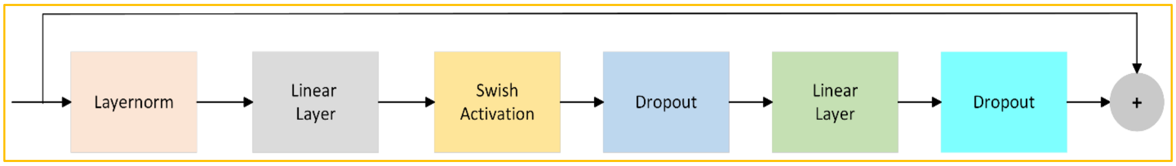

Fig. 4. Feed forward module.

CONFORMER BLOCK

As shown in Fig. 1. the conformer block is made of four structures containing the Convolution module and the Attention module which is sandwiched by two Feed Forward modules. Macaron-Net [29] gave this sandwiched idea and according to that two half-step feed-forward layers were introduced in the Transformer block instead of feed-forward layer, one before the attention layer and one after. If we define Mathematically, Conformer block $i\ for\ input\ x_i$, the block output yi is:

$$\tilde{x}_i = x_i + \frac{1}{2}FFN(x_i) \tag{2-1}$$

$$x_i^{'} = \tilde{x}_i + \text{MHSA}(\tilde{x}_i) \tag{2-2}$$

$$x_i^{''} = x_i^{'} + \text{Conv}(x_i^{'}) \tag{2-3}$$

$$y_i = \text{Layernorm}(x_i^{''} + \frac{1}{2}FFN(x_i^{''})) \tag{2-4}$$

FFN is Feed forward module, MHSA is Multi-Head Self-Attention module, and Conv is Convolution module.

This conformer structure having sandwiched Macaron-net style layers have a significant improvement than the single feedforward module.

DESIGN AND IMPLEMENTATION OF URDU SPEECH RECOGNITION FOR THYROID ULTRASOUND REPORT

The ultrasound report would be incomplete without a description of the thyroid ultrasound and a summary of the findings. Currently, ultrasound results are structured in a predetermined way. Every doctor describes the phenomena of ultrasound examination in his or her method, and there is no generally great standard for controlling it within the set format parameters. While manual input is time-consuming, this content requires it. Consider that a doctor's dictation is recognized by a speech input device that is fast and accurate. It is vital to create a medical lexicon and populate it with the most essential and regularly used terms. The algorithm searches the medical lexicon to find and match the terms that doctors use and then translates them into sentences. In Fig. 5, the Urdu speech recognition system for the Thyroid ultrasonography report consists essentially of three modules.

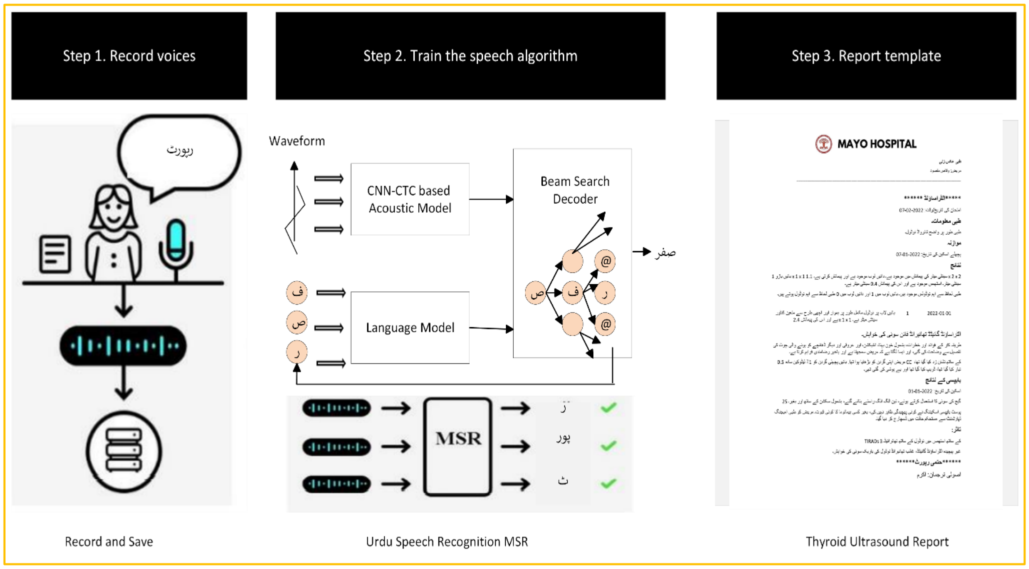

Fig. 5 Urdu speech recognition system based Thyroid ultrasound report

During the doctor's examination, the technology uses a microphone to gather voice data, which it then uses to synthesize digital speech. An Urdu speech recognition system then processes this input. As the last step, a report will be generated from the text you have created. Fourth, the patient will be given a copy of the examination report after the doctor checks it accurately. It will be printed.

SPEECH INPUT SYSTEM

The Urdu speech recognition system is linked to the ultrasound report's voice input module. It can automatically import patient and doctor data and bring it into the examination interface for review. The doctor selects a disease template based on the patient's voice input following a prompt from the system.

URDU SPEECH RECOGNITION SYSTEM

The basic role of the speech system is to translate information from speech into text. The recognition algorithm extracts the appropriate speech characteristics from the source speech. As network input for the speech characteristics, an acoustic model, a pronunciation dictionary, and a language model were employed. The network calculates the likelihood of each potential text information, and the text information with the highest probability is chosen for speech recognition. The conformer, acoustic and language models must be trained using the proper Speech databases and text sets.

REPORTING ON THYROID ULTRASOUND

The doctor can alter the ultrasound description and conclusion, collect and print the ultrasound image, and print out the ultrasound examination report using voice control. Inquiry into the past is the focus of this module. Any patient whom this system has examined can get information about their medical history and treatment from other patients who have been examined by it. With the correct information, doctors can make ultrasound diagnoses more quickly and accurately. Other similar patients' ultrasound descriptions may also be used to generate more accurate ultrasound examinations.

CTC-CNN, CTC-DCNN and Conformer models are developed in this chapter. Urdu speech recognition is constructed in this chapter, including the construction of the experimental platform and its components, composition and implementation of each component in the system, complete acoustic and conformer model training, and the Urdu speech recognition experiment.

II. EXPERIMENTAL PREPARATION

EXPERIMENTAL PLATFORM CONSTRUCTION

There are some of the most popular speech recognition frameworks which are being used today, such as Caffe, Kaldi, Torch, Tensorflow, Visual Studio Code and Theano. *Kaldi* is a framework that several researchers have used to investigate speech recognition. At the same time, Tensorflow was chosen to construct the Urdu speech recognition system in this thesis. Tensorflow is an open-source computing framework suggested by the Google team in 2015. Tensorflow simplifies the model building process by integrating several functional packages and encapsulating some complex functions. One of the most commonly used development tools for researchers is

Tensorflow. Most of the functions required by deep learning can be found in Tensorflow, a framework structure that can effectively interpret data flow graphs. Each side of the structure in Tensorflow depicts the data interaction between the operation nodes, with nodes representing various function operations. There are tensors, which represent the data communicated between nodes in a multi-dimensional matrix or vector, and flows, which indicate the flow of information through the diagram. This is a diagram of the Tensorflow architecture, as shown in Fig. 6.

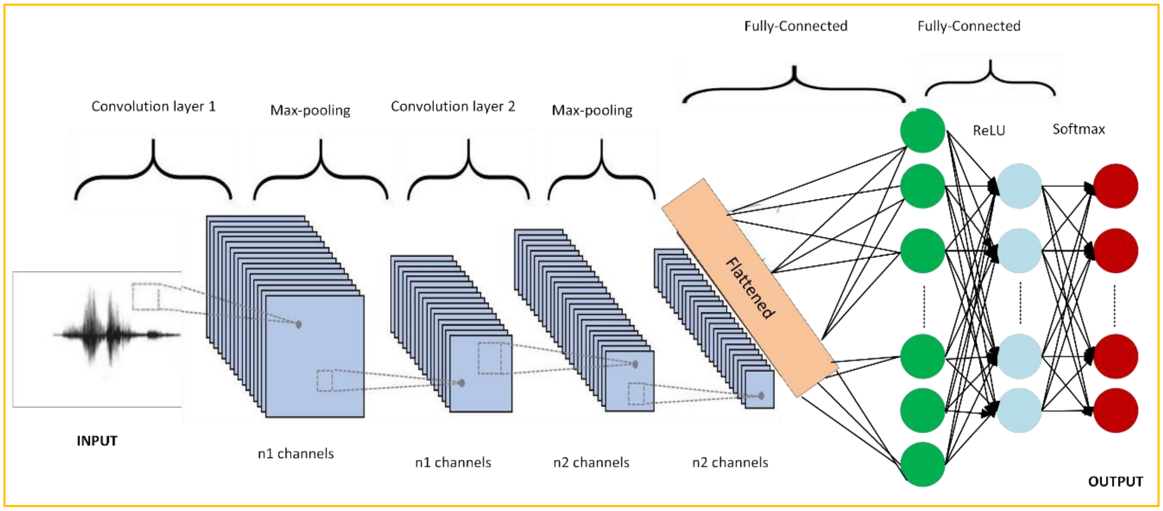

Fig. 6 The structural sketch of Tensorflow

Our experimental platform was built with windows, allowing us to simulate and test. The following are the steps:

(1) Make sure your computer's hardware is up to date. By utilizing the server, one can develop and implement a system and its functions. Many of them, including The CPU, an 8-core Intel i7-9750. Selecting an NVIDIA GTX1650 and 4000MB of VRAM for the GPU.

(2) Install Latest Python on Windows.

(3) Setup the Python requirements. The majority of the packages in this dependency list are dedicated to the extraction, processing, and calculating of speech. A wide range of tools may be found in these libraries, such as NumPy, Libros, and tflearn. Libros is used to extract features from voice data, among other things. Deep learning library tflearn provides Tensorflow with APIs to better construct experimental environments and conduct operations. The training data is preprocessed using Scikit-image. Once we have extracted the feature vector from the speech data, the speech feature can be treated as a map, just like an image. Data processing and analysis are performed with the help of Numpy and Pandas.

(4) Please make use of PyCharm by installing it on your computer. After selecting Tensorflow as the experimental framework, PyCharm was chosen as the compiler for the Urdu speech recognition system. In order to better exploit the coding and simulation capabilities of Python, the integrated development environment PyCharm was created.

DIALECT URDU SPEECH DATABASE

There was no Urdu speech database or text corpus to analyze. First, we developed an Urdu voice corpus with approximately 20000 utterances captured from 600 distinct sentences uttered by 10 male and female speakers. We separate the training and test sets in our database. The voice collection is more than 5 hours long in total. The training set is called the ABC group, and each group comprises 200 sentences; the test set is labeled the D group. The sampling frequency is set to 16000 Hz, and the sample size is set to 16 bits. The ABC group trains and learns the speech recognition model, while the D group represents the recognition effect. Table 1 below is the composition of our speech. The corpus consists of various thyroid medical reports. The reading atmosphere was a silent environment.

Table 1 The Dialect Urdu speech database

Data content	Training set	Test set
Speaker	10	4
Female	3	1
Male	7	3
Age Range (years)	20-35	19-25

Time Consuming (Hours)	5	1

Before using the database, data preparation is required, including the voice data required by our models, the voice model requires text data, and so on. The data structure is shown in Table 2. In Table 2, the word.txt contains information about the modeling unit, including the number of the modeling unit, and the corresponding content. The pronunciation dictionary information is stored in lexicon.txt in the format: phrase, syllable (including initials and tones). The text file contains voice information, including voice id and content. Wav.scp is stored in the wav folder for the specific voice and the corresponding path.

Table 2. Data content and format

File name	Information of content
Word.txt	B2_250 The maternity ward of the children's health center was equipped
Lexicon.txt	health care thyroid
Text	B2_250 Thyroid ultrasound contextual details
Wav.scp	C:\data\wav\train\A2_250.wav
Word.3gram.lm	-4.829394 thyroid -0.18121

EVALUATION

Word sequences are the most common output of continuous speech recognition. The recognition results are compared with the right annotation sequence using a dynamic programming approach. Insertion mistakes, deletion errors, and substitution errors are the three forms of errors that exist. Insertion mistakes occur when other words are inserted between two tags adjacent to each other, deletion errors occur when the corresponding words to a tag are not discovered in the results being recognized, and substitution errors occur when the words that are identified and associated tags are inconsistent. Assuming that a test set has N labels, the number of insertion errors is I, the number of deletion errors is D, and the number of errors is R. The evaluation index was the word error rate.

$$\mathrm{WER} = \frac{\mathrm{I} + \mathrm{D} + \mathrm{R}}{\mathrm{N}} * 100\% \qquad (3\text{-}1)$$

III. EXPERIMENTAL RESULTS AND ANALYSIS

EXPERIMENTAL RESULTS

In this section, we use tensor flow to design and build several URDU speech recognition systems based on CNN, CTC-CNN, CTC-DCNN, Conformer and other models and complete four groups of experiments. In these four experiments, firstly, the Convolutional neural network model of the shallow layer is analyzed and verified. Then the recognition accuracy of the proposed end-to-end deep Convolutional neural network acoustic modeling is analyzed. At the same time, optimize the acoustic model for CTC-CNN and CTC-DCNN and train under different iterations, and the recognition results are shown in Fig. 5. The superiority of the modified CTC-DCNN acoustic model in this paper is verified by these comparative experiments. Finally, a preliminary experiment and analysis of the model in a noisy environment are carried out.

Experiment 1: Urdu Speech recognition based on CTC-CNN and related CNN acoustic models. The experimental results are shown in Table 3. It can be seen that under the same iteration times, the recognition effect of the CTC-CNN acoustic model proposed in this paper is better than that of the CNN acoustic model. In this section, through Tensorflow, Python is used to build the Urdu Speech recognition system and compile the code. After the Urdu Speech data training set, the test set speech input recognition system, after feature extraction, complete pattern matching, and finally get the corresponding output.

Table 3 Recognition results of systems

Model	**Convolutional layer activation function**	**Full connection layer activation function**	**WER%**
CNN	sigmoid	ReLU	16.9%

CNN (maxout)	maxout	maxout	16.5%
CTC-CNN	sigmoid	ReLU	15.7%
CTC-CNN (maxout)	Maxo	maxout	15.3%

Experiment 2: The recognition is based on CTC-DCNN and other models. It can be seen that the recognition effect of the CTC-DCNN model is 12.9%, while the error rate of UDRU speech words is reduced to 12.2% under the improved CTCDCNN acoustic model of maxout function.

Table 4 Recognition results of systems

Model	**Convolutional layer activation function**	**Full connection layer activation function**	**WER%**
DCNN	sigmoid	ReLU	14.2%
DCNN (ReLU)	ReLU	ReLU	13.8%
CTC-DCNN	ReLU	ReLU	12.9%
CTC-DCNN (maxout)	maxout	maxout	12.2%

Experiment 3: The recognition is based on Conformer and other models. It can be seen that under the same iteration times, the recognition effect of the conformer acoustic model proposed in this paper is better than that of the CNN acoustic model.

Table 5 Recognition results of systems.

Model	**WER%**
CTC-CNN (maxout)	15.3%
CTC-DCNN (maxout)	12.2%
Conformer	8.8%

*Experiment 4***:** The effect of different iterations on the model. After the previous model training, it is found that the Conformer model has the highest accuracy of speech recognition when the number of selection iterations is 16 times. In order to verify whether there are more suitable iterations, this paper is for the Conformer, CNN model, DCNN model, CTC-CNN model, CTC-DCNN model, and maxout improved CTCDCNN model were retrained, and the recognition effect was verified under different iterations as shown in Figure 7.

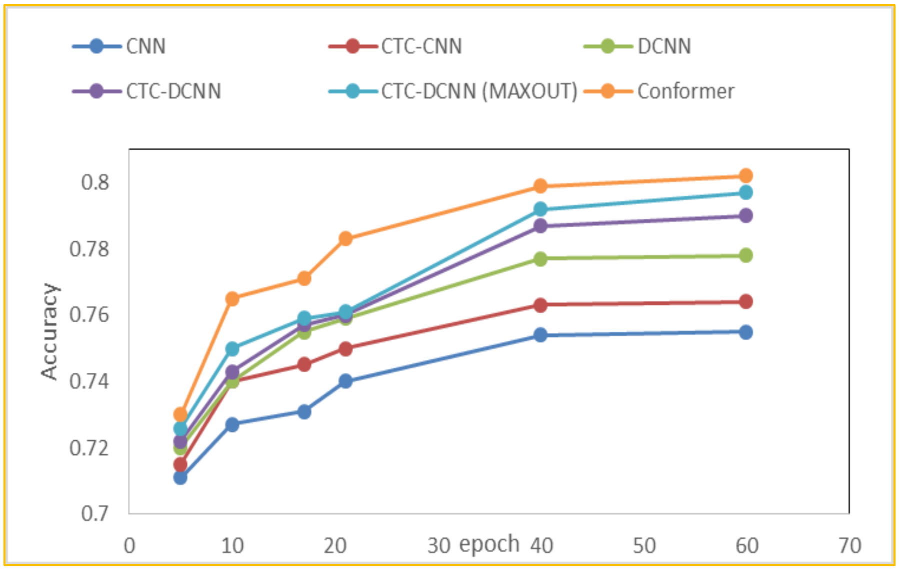

Fig. 7 Effect of iteration number

Experiment 5: The effect of noise on the Urdu Speech recognition system. In this section, after analyzing the Conformer model, in order to explore whether the model can still achieve a good recognition effect under a noisy environment, the model in the noisy environment is trained, and the recognition result is obtained through the test set. This model did a preliminary study with or without noise. The result of the recognition is shown in Figure 8.

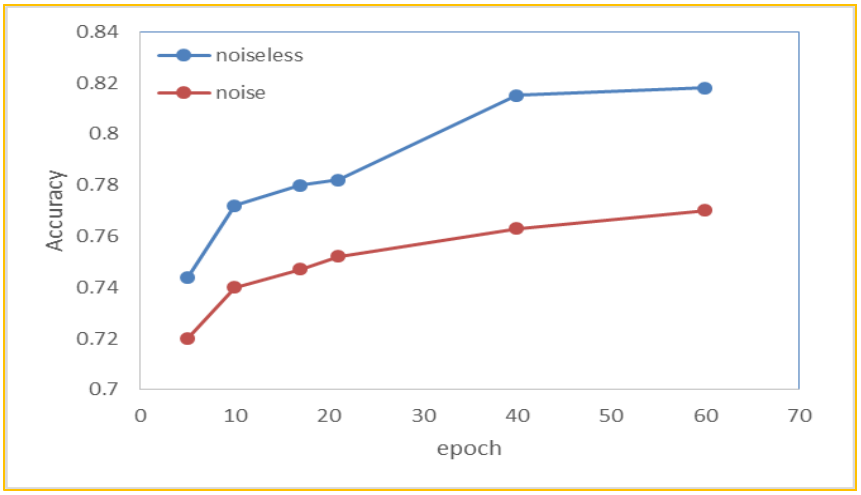

Fig. 8 The Effect of Noise on Recognition

ANALYSIS OF EXPERIMENTAL RESULTS

From Table 3, different acoustic models are selected, and the effect of Urdu Speech recognition is also different. In Urdu Speech recognition, the homonym, synonym, and other factors existing in Urdu speech will further aggravate the difficulty of the recognition process, so the recognition effect has not been particularly ideal. In the CNN acoustic model, the word error rate of the test is 16.9%; if the maxout function is employed as the CNN activation function, after training and testing again, the recognition impact improves by 0.4 percent. After the end-to-end structure design of the CTC-CNN acoustic model, it is found that the CTC structure can directly optimize the input and output sequence, which is closer to the distribution of real speech and more "intelligent" in the recognition process, making the robustness of the model further improved, and the recognition effect has nearly 1.2% improvement compared with CNN model. Compared with CNN and CNN (maxout) model, CTC-CNN, and CTC-CNN(maxout) model, it is found that the recognition effect is better when the activation function is maxout. This is because the maxout function has a strong fitting ability. Given

http://dx.doi.org/10.29322/IJSRP.12.07.2022.p12703

a constant gradient, it can effectively improve the gradient disappearance phenomenon in the model during the convolutional process to improve the speech effect.

Table 4 shows that different deep convolutional neural network models produce different results in speech recognition, as demonstrated by the findings. The error rate of the DCNN model for word recognition is 14.2 percent when the Convolutional and pooling layers are alternately connected. The DCNN (ReLU), CTC-DCNN model, and CTC-DCNN(maxout) models are built using the residual structure of the DCNN model. Speech recognition is accurate to a level of 13.8%, 12.9%, and 12.2%, respectively. Residual structure (ReLU) can be better conserved and improved as compared to the DCNN model. At the same time, the residual structure itself can improve the gradient, so it can get a better recognition effect. Compared with the DCNN(ReLU) model, the CTC-DCNN model improves the recognition accuracy by 0.9%, which shows that in the deep convolutional structure, CTC can still play a good role and improve the robustness of the model.

Compared with Table 3 and Table 4, it can be seen from the experimental data that the speech recognition accuracy of the deep convolutional network is improved by 0.9% to 4.7% compared with the shallow convolutional network, which shows that with the deepening of the network layer, the speech feature can be extracted better. But Table 5 clearly shows the dominance of conformer model in terms of its word error rate accuracy. Conformer gives 8.8% word error rate accuracy while CTC-CNN(maxout) and CTC-DCNN(maxout) gives 15.3% and 12.2% respectivel which clearly shows the outperformance of conformer model.

It can be found in Fig. 7 that the accuracy of the conformer and all the acoustic models increases with the number of iterations. This is because, with the increase in training times, the model can learn more accurate speech features. In this section, the number of iterations is selected as 5 at the beginning, and the test results are not very ideal. When the iteration number is increased and the test is conducted, it is found that the accuracy of the model has a significant improvement and then gradually tends to be stable. The increase in the number of iterations means that the training time of the model is longer. When the Iteration number exceeds 40, the results tend to be stable gradually, reaching a "saturation value." At this time, the model for Urdu Speech recognition reaches a more ideal value, and the learning effect is better.

Fig. 8 shows the results of the Conformer model for Urdu Speech recognition in a noisy environment. As shown in Fig. 8 that the recognition effect of the Conformer model in a quiet environment is significantly better than that in a noisy environment. Through the preliminary analysis of the accuracy of the noisy environment, it can be seen that the recognition rate of the conformer model in a noisy environment is significantly lower than that in a quiet environment, and the recognition effect still needs to be improved, which is also the focus of our next work.

IV. CONCLUSION

Image and speech recognition are two of the most common uses of deep learning CNN network topology. The first half of this thesis was devoted to the framework and basic principles of speech recognition, as well as the major state of thyroid ultrasonography report creation. Analysis of speech recognition and feature extraction in the Urdu Speech recognition system was conducted. As we can see, the acoustic model is a critical component of a speech recognition system but Conformer presents the best results. Conformer has a direct impact on speech recognition accuracy. After that, we concentrated on the speech recognition method with different acoustic models as well as conformer.

End-to-end Convolutional neural network-based acoustic modeling. On the one hand, when CNN is used for speech recognition, it can guarantee the invariance of the speech signal in time and space. The model's resilience will be further enhanced by reducing the number of parameters. A CTC-CNN acoustic model for Urdu Speech recognition was designed and established. The classical acoustic model is compared to CTC and a Convolutional neural network. A more accurate word string can be found by using an end-to-end structure, which does not need labeling data or performing other actions on it. For Urdu Speech recognition, the CTC-CNN model developed in this research outperforms the standard CNN model. The CTC- CNN acoustic model's error rate for Urdu speech recognition is 17.7 percent, which is lower than the CNN model's error rate by 1.2 percent. Deeper layers of the network suffer from the gradient disappearance phenomena, which affects the recognition results because the parameters of the shallow Convolutional network training are huge. The residual block is used in the Convolutional neural network structure to generate a more accurate audio model in this study. Also, we developed a new CTC-DCNN model, where the optimization of this model was performed by employing the maxout function. A new and enhanced Urdu speech recognition system has been developed and built. To further improve gradient disappearance, a maxout function was introduced to the acoustic model developed in this research. The model's end-to-end structure was also found to match the entire speech sequence. The error rate of the Urdu Speech recognition system is 14.2 percent after training and testing the Urdu Speech database. A 2 percent reduction in word mistake rates is achieved over the DCNN acoustic model.

There have been experiments on various Urdu speech recognition systems, and the residual block structure is better at retaining Urdu speech characteristics in the deep Convolutional neural network, and the end-to-end structure analyses the probability output of the entire Urdu speech, while the maxout function improves the gradient problem in the network, which ensures the recognition effect of the end-to-end deep Convolutional neural network.

Undoubtedly, conformer-based end-to-end automated speech recognition has gotten a lot of attention since it outperforms recurrent neural network-based versions and that's exactly what happened in our research. Conformer parallel computing is more efficient than recurrent neural networks, because the conformer-based end-to-end automatic speech recognition model was trained on very limited medical lexicon. There is a lot of work still required to make it robust for thorough medical terminologies and all the reports. The model may also perform poorly with accented speech.

At last thyroid ultrasound report generation based on Urdu, Speech recognition is possible as long as we have an efficient speech recognition system for Urdu Speech-language. As we find out that our purposed thyroid ultrasound report generation based on Urdu Speech recognition can be built up by including three significant modules: speech input system, Urdu Speech recognition system, and thyroid ultrasound report. The system searches and matches the contents dictated by doctors in the vocabulary and converts them into words.

In this work we had to deal with some of the chalanges that we did not know. We trained different models including various acoustic variences and conformer on Urdu langauge. This language structure is very different to most commonly spoken langauges. The data set arranged in this work was only from the doctors because patients with thyroid probelms could not properly comprehend our research. Most of the patients, due to lack of trust, were not comfortable to give us their details because of technophobia. Our country does not have centralized hospital system. Everything is independent and if there is an organization who collects data from different hospitals, their data security and management is not upto the mark. So, data collaboration is a highlighted issue as there is no database supported by the government, so everytime one who wants to carry out the research must collect the data by himself. The versatility of data is another problem, as the society lacks the awareness, collecting data from different areas to get versatile data was problematic. The model, Conformer, we used along with other accoustic models is a latest model. It requires latest machinery to run these models and the technology is not that updated in Pakistan.

The future of speech recognition technology beholds efficient Conformer models. Techniques such as quantization and weights pruning to reduce inference time will become the prerequiste in model structuring. Adaptive fusion mechanism, that weighs every modality on the basis of noise level, would be very interesting to further investigate in the future. In the future, Conformer training on various dialects of Urdu and other local languages have to be employed to make in order to implement this model in all regions of Pakistan. A great extent of research work has to be done to further improve Conformer accuracy while dealing with all the accents of Urdu speech recognition. Moreover, why do we have to stick ourselved to only thyroid ultrasound report generation, when we can expand this to all the medical fields. Currently, Urdu speech recognition technology is infant and when it couples with medical lexicon, the horizons are unthinkable. Furthermore, end-to-end automated speech recognition is gaining traction in academics and industry. The acoustic model, pronunciation model, and language model are all combined into a single neural network. It produces competitive performance with Urdu speech recognition while also streamlining the training and decoding processes.

REFERENCES

[1] S.A. Lowe, Diagnostic radiography in pregnancy: Risks and reality, Aust. New Zeal. J. Obstet. Gynaecol. (2004). https://doi.org/10.1111/j.1479-828X.2004.00212.x.

[2] S.K. Hasnain, M.S. Awan, Recognizing spoken Urdu numbers using fourier descriptor and neural networks with Matlab, 2008 Second Int. Conf. Electr. Eng. (2008) 1–6.

[3] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech., J. Acoust. Soc. Am. 87 4 (1990) 1738–1752.

[4] C.C. Chiu, T.N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R.J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, M. Bacchiani, State-of-the-Art Speech Recognition with Sequence-to-Sequence Models, in: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., 2018. https://doi.org/10.1109/ICASSP.2018.8462105.

[5] K. Rao, H. Sak, R. Prabhavalkar, Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer, in: 2017 IEEE Autom. Speech Recognit. Underst. Work. ASRU 2017 - Proc., 2018. https://doi.org/10.1109/ASRU.2017.8268935.

[6] H. Sak, A. Senior, F. Beaufays, Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition, (2014).

[7] X. Li, X. Wu, Long short-term memory based convolutional recurrent neural networks for large vocabulary speech recognition, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. 2015–Janua (2015) 3219–3223. https://doi.org/10.21437/interspeech.2015-648.

[8] H. Sarfraz, S. Hussain, R. Bokhari, A.A. Raza, I. Ullah, S. Pervez, A. Mustafa, I. Javed, R. Parveen, Speech Corpus Development for a Speaker Independent Spontaneous Urdu Speech Recognition System, Proc. O-COCOSDA 2010. (2010) 1–6.

[9] N. Yalta, S. Watanabe, T. Hori, K. Nakadai, T. Ogata, CNN-based multichannel end-to-end speech recognition for everyday home environments, Eur. Signal Process. Conf. 2019–Septe (2019). https://doi.org/10.23919/EUSIPCO.2019.8902524.

[10] S.H. Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, Lite Transformer with Long-Short Range Attention, arXiv:2004.11886. (2020).

[11] H. Sarfraz, S. Hussain, R. Bokhari, A.A. Raza, I. Ullah, Z. Sarfraz, S. Pervez, A. Mustafa, I. Javed, R. Parveen, Large vocabulary continuous speech recognition for Urdu, in: Proc. 8th Int. Conf. Front. Inf. Technol. FIT'10, 2010. https://doi.org/10.1145/1943628.1943629.

[12] A.W. Yu, D. Dohan, M.T. Luong, R. Zhao, K. Chen, M. Norouzi, Q. V. Le, QaNet: Combining local convolution with global self-attention for reading comprehension, in: 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., 2018.

[13] M.U. Farooq, F. Adeeba, S. Rauf, S. Hussain, Improving large vocabulary Urdu speech recognition system using deep neural networks, in: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2019. https://doi.org/10.21437/Interspeech.2019-2629.

[14] A. Graves, Sequence transduction with recurrent neural networks, arXiv Prepr. arXiv1211.3711, 2012. (2012).

[15] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, S. Kumar, Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss, in: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., 2020. https://doi.org/10.1109/ICASSP40776.2020.9053896.

[16] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J.M. Cohen, H. Nguyen, R.T. Gadde, Jasper: An end-to-end convolutional neural acoustic model, in: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2019. https://doi.org/10.21437/Interspeech.2019-1819.

[17] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, Y. Zhang, Quartznet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions, in: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., 2020. https://doi.org/10.1109/ICASSP40776.2020.9053889.

[18] W. Han, Z. Zhang, Y. Zhang, J. Yu, C.C. Chiu, J. Qin, A. Gulati, R. Pang, Y. Wu, ContextNet: Improving convolutional neural networks for automatic speech recognition with global context, in: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2020. https://doi.org/10.21437/Interspeech.2020-2059.

[19] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-Excitation Networks, IEEE Trans. Pattern Anal. Mach. Intell. (2020). https://doi.org/10.1109/TPAMI.2019.2913372.

[20] I. Bello, B. Zoph, Q. Le, A. Vaswani, J. Shlens, Attention augmented convolutional networks, in: Proc. IEEE Int. Conf. Comput. Vis., 2019. https://doi.org/10.1109/ICCV.2019.00338.

[21] B. Yang, L. Wang, D.F. Wong, L.S. Chao, Z. Tu, Convolutional self-attention networks, in: NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., 2019. https://doi.org/10.18653/v1/n19-1407.

[22] A. Gulati, J. Qin, C.C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-augmented transformer for speech recognition, in: Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2020. https://doi.org/10.21437/Interspeech.2020-3015.

[23] S. Karita, X. Wang, S. Watanabe, T. Yoshimura, W. Zhang, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N.E.Y. Soplin, R. Yamamoto, A Comparative Study on Transformer vs RNN in Speech Applications, in: 2019 IEEE Autom. Speech Recognit. Underst. Work. ASRU 2019 - Proc., 2019. https://doi.org/10.1109/ASRU46091.2019.9003750.

[24] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, R. Salakhutdinov, Transformer-XL: Attentive language models beyond a fixed-length context, in: ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., 2020. https://doi.org/10.18653/v1/p19-1285.

[25] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D.F. Wong, L.S. Chao, Learning deep transformer models for machine translation, in: ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., 2020. https://doi.org/10.18653/v1/p19-1176.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Adv. Neural Inf. Process. Syst., 2017.

[27] L. Dong, S. Xu, B. Xu, Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition, in: ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., 2018. https://doi.org/10.1109/ICASSP.2018.8462506.

[28] Q.V. Le Prajit Ramachandran, Barret Zoph, Searching for Activation Functions, arXiv:1710.05941. (2017).

[29] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, T.-Y. Liu, Understanding and Improving Transformer From a Multi-Particle Dynamic System Point of View, (2019) 1–15.

AUTHORS

First Author – Naeem Haider, Mater of Philosophy, School of Information Science and Engineering, Xinjiang University, China, naeemhaider0007@gmail.com

Second Author - Wei Fuyuan, School of Information Science and Engineering, Xinjiang University, China

Correspondence Author – Wenzhong Yang, xwz_xy@163.com