DEEP RECURRENT NEURAL NETWORK BASED AUDIO SPEECH RECOGNITION SYSTEM

Savitha G[1], Shankar Gowda B N[2] , Shashwat Shahi[3]
[1]Associate Professor, Department of Computer Science and Engineering, RVITM, Bangalore-560076, India
[2]Associate Professor, Department of Computer Science and Engineering, BIT, Bangalore-560004, India
[3]Department of Computer Science and Engineering, SaIT, Bangalore-560097, India
savithamadhusudan@gmail.com, shashwats538@gmail.com

ABSTRACT:

Speech recognition system has become an integral part of how the computer technologies can be used to influence and improve the human work and activities. These are being used extensively, from personal assistants to self-driving car Human Computer Interfaces (HCIs) and other industries as well. While the most common approach to speech recognition system building is using the Hidden Markov Models (HMMs). The HMM models assume a specific structure of the data and unable to capture temporal dependencies. This paper however presents a unique approach for isolated word recognition based on deep learning models using Recurrent Neural Networks (RNNs) particularly, which can perform end to end speech recognition without any assumption of structure in data using Bidirectional LSTM (BiLSTM). The network proposed in the paper can learn both features in the data and capture temporal dependencies.

Keywords: Deep Learning, DeepSpeech, Recurrent Neural Networks (RNNs), Hidden Markov Models (HMMs), Isolated Word Recognition, BiLSTM.

1. INTRODUCTION

Speech Recognition System can be basically understood as a system in which a machine transforms the spoken language into text. These types of speech recognition systems are now a days used extensively in various industries. One of the most popular application of this system is the voice typing feature in many word processors and editors such as google docs etc. Designing and building a precise speech recognition system is considered as a complex task as it has to consider several sources of variation in data. These variations in the data can be the variation in speakers, the size of the vocabulary, the ambient noises, the accent of the speakers, the characteristics of the speaker such as age, gender etc. and so on. For example, a word "Apple" can be said by different people in different styles and the speech recognition system should detect all the cases and provide a valid output.

The most common approach to design a speech recognition uses the Hidden Markov Models (HMMs). This is because the speech data or the temporal data can be considered as a stochastic or a probabilistic process and for a short time slices of the data can be considered to be as independent of time. The HMM models can be trained on short slices of the speech data that represents a phoneme or a word. It then predicts the next phoneme or the next word that can be combined to form the speech text. One additional benefit of the HMM based models are that it can be trained even on less data and it gives comparatively higher accuracy than other neural network-based models when trained on less data. The major drawbacks of HMMs are that they often contain many unstructured parameters. Also, they cannot express dependencies between the hidden states. Apart from HMM, there are other neural network based deep learning models too, such as RNN which can perform end to end speech recognition.

Speech recognition can further be classified as two separate individual tasks i.e. Continuous Speech Recognition and Isolated Word Recognition. Isolated word recognition refers to transcribing individual words spoken with a clear delineation. On the other hand, continuous speech recognition refers to the way we normally speak sentences, recognizing them and then transcribing them into text. The isolated words

ISSN (Print): 2204-0595
ISSN(Online):22031731

are often spoken with clear pronunciation and it is easier to identify the boundaries while in the continuous speech, we can have a lot of variations in pronunciations, gaps and inter-word dependencies. The HMM models have been and can be used for both tasks but the HMM models assume a specific structure of the data. However, there are some recent models which can perform end to end speech recognition without any assumption of structure in data and the network can learn both the features in the data and capture temporal dependencies. The only requirement of these models is that they need more data to get trained than the HMM models.

2. RELATED WORKS

Alex Krizhevsky et al., [1] proposed that the Deep Neural Networks have been used successfully in many applications. One of the most important application of the deep neural network has been classification (i.e., mapping a fixed length vector to an output category). In case of structured problems, where mapping must be done between one variable length sequence to another variable length sequence, the neural networks alone are not sufficient for the job. Hence, they must be combined with some other sequential models such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs). One major drawback of this combined approach is that the models cannot be easily trained end-to-end and they make assumptions about the probability distribution of the data and about the structure of the data.

Ilya Sutskever et al., [2] illustrated a work in which sequence to sequence learning is basically a framework that addresses the problem of learning variable length input and output sequences. It basically works on two major components-An Encoder RNN and a Decoder RNN. The encoder RNN is used to map the sequential variable length input into a fixed length vector. Then, the decoder RNN then uses this vector and produces the variable length output sequence which is produced as one token at a time. It takes the input as the ground truth labels to the decoder during the training period. The model then performs a beam search and generates suitable candidates for the next step predictions during the inference. Attention based mechanisms can be used to improve these types of models significantly.

Minh-Thang Luong et al., [3] researched about the various diverse applications of the sequence-to-sequence frameworks in the field of image captioning, parsing, machine translation etc. Thus, having such diverse applications of this framework, it was also suggested that speech recognition too can be one of the direct applications of this framework.

Geetha K et al., [4] presented that the phonemes or the words can be used in designing speech recognition systems. A phoneme can be coined as the smallest unit of a language. The phonemes depend on the context. Speech Segmentation is the phase of identifying the end points of the acoustic units of the speech signal. The methodology that has been used for speech segmentation are preprocessing and Feature Extraction with Phoneme Boundary Detection and Spectral Transition Measure (STM). But this methodology leads to some disadvantages as well, like, low accuracy obtained and large phoneme detection.

D. Dhanashri et al., [5] explored that most of the speech recognition systems are currently based on the Hidden Markov Model (HMMs) whose main benefits are that it is a statistical framework that supports both temporal and acoustic modeling. While there are Neural Networks that can learn complex functions, can avoid many assumptions, support parallelism, can tolerate noise, and generalize effectively.

Hadi Veisi et al., [6] explained that the Artificial Neural Network (ANN) and Hidden Markov Models (HMM) are the most common methods for developing a speech recognition system. Improving the efficiency and accuracy of these systems is the major problem. To address this problem, they extracted the features of the speech signal using a combination of Deep Belief Network (DBN) and Connectionist Temporal Classification (CTC) output layer with Deep Bidirectional Long Short-Term Memory (DBLSTM) to create an Acoustic Model (AM) on the speech data set. They observed that when compared to the unidirectional models, the use of the bidirectional network improves the accuracy of the model and as

ISSN (Print): 2204-0595
ISSN(Online):22031731

compared to the Shallow Networks, the use of Deep Neural Networks improved the results.

Samuel Kriman et al., [7] developed a new end-to-end neural acoustic model designed in the form of multiple blocks for automatic speech recognition. The blocks had residual connections between them. This new model was based on deep neural network (DNN) along with 1D time-channel separable convolutional layers and was trained with Loss of Connectionist Temporal Classification (CTC). The model had lesser parameters than the other competing models and achieved a near state-of-the-art accuracy on Wall Street Journal and LibriSpeech.

Dinesh Kumar Vishwakarma et al., [8] researched about the implications of deep learning architectures. They provided a comparison of different deep learning architectures where they presented a summary of popular datasets, the key features of the dataset and the accuracy of different deep learning techniques. They reached to the conclusion that the different deep learning architectures can be used to solve complex problems like sentimental analysis.

Tomohiro Tanaka et al., [9] proposed to revise the hypotheses of Automatic Speech Recognition Systems based on hybrid of Deep Neural Networks and Hidden Markov Models (DNN-HMM). The Neural Network based end-to-end ASR Systems do not introduce heuristic problems and are known to provide comparative better performance. Sometimes the architecture suffers from problems like omission of important words in text generation phases and redundant generation. Therefore, the authors proposed a new architecture for rescoring hypotheses based on the neural speech-to-text language models (NS2TLMs). They demonstrated that because of the quality that NS2TLMs can handle short-duration utterances correctly they can improve the DNN-HMM hybrid ASR systems.

An Audio Video Speech Recognition (AVSR) was presented by Kuniaki Noda et al., [10] which can be highly efficient for audio corrupted by noise. A connectionist-hidden markov model (HMM) system was introduced for noise-robust AVSR. For acquiring noise-robust audio features, a deep denoising autoencoder was used and a Convolutional Neural Network (CNN) was used to extract the visual features. For integrating the acquired audio and visual HMMs, they suggested to use a multi-stream HMM (MSHMM) which are trained with their respective features independently. They studied a comparative analysis of the normal and denoised Mel-frequency Cepstral Coefficients (MFCCs). The analysis revealed that the denoised MFCCs gave better results than the normal MFCCs.

Mohammad Mehdi Homayounpour et al., [11] demonstrated a new method to further reduce the recognition error rate using the Adaptive windows convolutional neural network (AWCNN). The variation in joint temporal-spectral features was analyzed by AWCNN. The analysis of this variation makes the model more robust against both types of speaker variation i.e. Both inter-speaker variation and intra-speaker variation. A new residual learning was proposed by the authors which will help in utilizing the information present in deep layers more efficiently. They found that their proposed architecture, in comparison with other state-of-art methods reduced the absolute error rate by 7%.

Chengyi Wang et al., [12] explained in their work that in offline mode, the attention-based Transformer models have achieved significant results but the Transformer model in the streaming mode, to maintain its recognition accuracy, usually incurs significant latency. To tackle this problem, for transformer models the authors proposed a novel low latency streaming approach, that is composed of a recognition network and a scout network. In this architecture, the scout network, without seeing any future frames, predicted the whole boundary of the word and the recognition network is used by utilizing the information from all the frames, on the next sub word, before the predicted boundary.

Soufane Hourri et al., [13] demonstrated a work in which the authors proposed a unique way of speaker recognition using deep neural networks (DNNs), where they used DNN to learn the distribution of features. The enhanced feature vectors were formed from the transformation of extracted feature vectors. The authors claimed that in both clean and

ISSN (Print): 2204-0595
ISSN(Online):22031731

noisy conditions the proposed architecture outperformed i-vector.

Hu Hu et al., [14] developed a new architecture to improve the RNN Transducer based modeling for speech recognition. The architecture proposed to have improved RNN training in multiple aspects. The RNN-T training algorithm was optimized by the authors, which resulted to have reduced the memory consumption with faster training speed. They also proposed better model structures which resulted in improved accuracy with smaller footprint.

Xiaoyu Qiu et al., [15] presented a work on Speech and Gesture Recognition using Multimodal Fusion architecture which is based on Leap motion-based hand motion recognition and microphone-based speech command recognition. To process the instructions of each single-mode input, the proposed architecture used different deep neural networks. It then uses techniques like similarity comparison and keyword retrieval to detect the correctness of the recognition command. The speech recognition uses the keywords of the gesture recognition. Then the words of speech recognition are processed using word segmentation, whose results are compared with the keywords of gesture recognition. Then the two comparison results are merged to find the final command.

3. METHODOLOGY

3.1. Data Source

The dataset used here was obtained from Kaggle. The dataset that we have used here is originally belongs to speech recordings of the **Linguistic Data Consortium (LDC).** While the original dataset of the LDC is quite huge (several gigabytes), the dataset that we have obtained from the Kaggle and used here in the project is a subset of the original dataset. The dataset that we have taken contains speech recordings of 16 different speakers from 8 different dialect regions. The data set also has one male and one female speaker from each dialect region. This data set contains the speech recordings in .wav format. The audio data is of single channel with 16 kHz sampling and 16-bit sample with PCM encoding. We have also used another dataset i.e., Free Spoken Digit Datasets (FSDD) that we have obtained from GitHub. FSDD is simple audio/speech dataset consisting of recordings of spoken digits in .wav files at 8kHz. The recordings are trimmed so that they have near minimal silence at the beginnings and ends. This dataset has over 1000+ recordings of 6 speakers of English accent.

3.2. Data Preprocessing

The data preprocessing is one of the most important phases in designing a Speech Recognition System.

To preprocess the data, we have transformed the raw audio data into its frequency spectrum. The frequency spectrum, also called as the power spectrum is like a fingerprint for the data in which the raw audio is broken into constituent parts or frequencies. This representation of the audio data helps in identifying that which are the frequencies that dominate over the others in the signal.

For preprocessing the data, we have used the following python libraries:

3.2.1. NumPy

NumPy is the common abbreviation used for Numerical Python. It is a perfect tool that is being used for scientific computing and performing basic and advanced array operations. Python uses different library features to operate on matrix and arrays. It helps in preparing exhibits that can store estimations of the indistinguishable information type and make performing math procedures on clusters and their vectorization simpler. Also, it increases the performance which is achieved by vectorizing mathematical operations and accelerating the execution time.

3.2.2 Tensorflow

TensorFlow is an open-source library which was developed by Google primarily for deep learning applications. It also supports traditional machine learning. Originally, TensorFlow was developed for large numerical computations. However, later it proved to be extremely useful for deep learning development as well.

3.2.3 Librosa

Librosa is a python package commonly used for audio data analysis. It provides all the building blocks necessary to create audio information retrieval systems.

After the preprocessing of the data, the dataset was divided into two parts for training and testing. For the

ISSN (Print): 2204-0595
ISSN(Online):22031731

training purpose, we have used 90% of the available data. And for the testing process, we have used the remaining 10% of the dataset.

3.3. Feature Extraction

For feature extraction of the data, we have used the **Mel frequency cepstral coefficient feature (MFCC)** of the audio signal. MFCC is kind of a frequency spectrum or power spectrum. It is obtained from short time frames of the signal. It is based on the assumption that for short duration of time, like the order of 20ms to 40ms, the frequency spectrum does not change much. Therefore, based on the above assumption we sliced the signal into such short frames and then the spectrum is computed for each slice. We have used the librosa library to perform this task and extract the MFCC feature.

After extracting the MFCC features, a dummy context was added to the front and the back of each time slice using the functions of NumPy library. The MFCC features were returned with the past and future context. These MFCC features were recursively extracted from each of the .wav file and the corresponding transcribed text was read from the text file. Then, the Non–Alphabetic Characters were removed from the transcribed text. Finally, the raw text was converted to integer labels that were used as the target labels for model training.

4. ARCHITECTURE

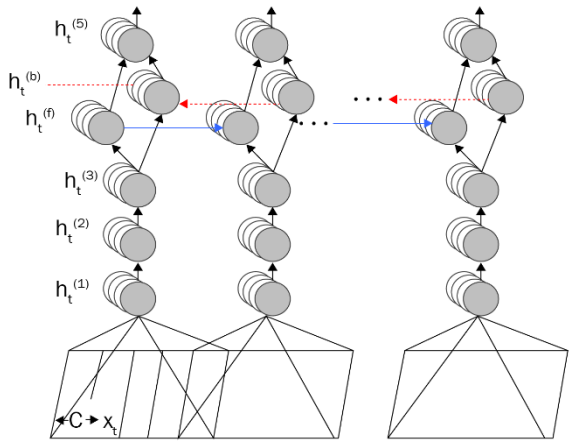

Figure.1. The architecture of the model

Figure.1. shows the deepspeech's model architecture used in this paper. It shows the hidden layers with the blue arrows (solid line) and the recurrent layer denoted by the red arrows (dotted line). The diagram also shows the audio input along with the time sliced MFCC features. In the diagram, X_t refers to the time sliced MFCC features at time t.

We are proposing to use the Deep Speech Architecture [16] for designing the speech recognition systems. The DeepSpeech is a state-of-the-art speech recognition system developed using end-to-end deep learning. It can also be understood as an architecture where deep learning replaces traditional hand engineered speech to text algorithms. The DeepSpeech architecture is simpler than the traditional architectures that relied on an extremely complicated network of processing pipelines. Also, when used in noisy environments the performance of traditional systems was poor while the DeepSpeech architecture does not need any special components to model speaker variation, reverberation, or background noise. The DeepSpeech model performs well, independent of speaker adaptation as it directly learns from the data.

The DeepSpeech model's architecture consists of a stack of fully connected hidden layers. The stack is followed by a bidirectional RNN. It also contains additional hidden layers at the output. The stack consists of a total of 5 layers. The first 3 layers are the Non-Recurrent layers. These act like a preprocessing step to the RNN layer. The output of these layers is passed through an additional clipped Rectified Linear Unit (ReLUs) to prevent the activations from exploding. The input audio feature is the Mel frequency cepstrum coefficients (MFCC) that are sent to the nonrecurrent layers in time slices of spectrograms. The spectrum data also contains the past and the future contexts in addition to the usual time slices. The fourth layer is the RNN layer. Both the forward and the backward recurrence is present in this layer. The fifth layer takes the concatenated outputs from both the forward and backward recurrence. This layer then produces an output which is finally fed to the final softmax layer, that predicts the character probabilities.

The character probabilities are calculated at each time slice t and character k in the word using:

ISSN (Print): 2204-0595
ISSN(Online):22031731

$$h_{t,k} = \hat{y}_{t,k} \equiv \mathbb{P}(c_t = k|x) = \frac{\exp(W_k \ h_t \ + b_k \ )}{\sum_j \exp(W_j \ h_t \ + b_j \ )}. \quad (1)$$

Here, b_k and W_k denote the k'th bias of the final softmax layer and k'th column of the weight matrix respectively, and h_t denotes the value computed for the layer previous to the softmax layer.

5. ALGORITHM

In the proposed model, we have tried to use and implement the original deepspeech architecture. The model consists of both the recurrent and non-recurrent layers. The first three layers of the model are non-recurrent. We specify the weights of the layers are H1, H2 and H3 and the biases of the layers are B1, B2 and B3. The outputs of these layers are passed through a clipped ReLU function. This ReLU functions helps in avoiding of exploding activations. The first hidden layer's weights have shape of:

n_inp + 2*n_inp*n_ctx, n_h (2)

Here, n_h denotes the number of hidden units which we have set as 1024, n_inp denotes the MFCC features and n_ctx denotes context which we have used as 9. Also, the shape of the weight of this layer is the same as MFCC input with context. For each time t, the first 3 layers are computed by:

$$\mathbf{h_t^{(l)} = g\ (W^{(l)}\ h_t\ ^{(l-1)} + b^{(l)})} \quad (3)$$

Here, g(z) = min {max {0, z}, 20} is the clipped rectified-linear (ReLu) activation function. $b^{(l)}$, $W^{(l)}$ are the bias parameters and weight matrix for layer l.

The recurrent layer is bidirectional LSTM with dropout. The concatenated output of the forward and backward LSTM is input to the next hidden layer.The final two hidden layers have weights as H4 and H5 and biases as B4 and B5. The fourth layer has the output ReLU activations clipped with dropout. Then, the fifth layer finally outputs the probabilities for number of alphabets plus blank, one character at a time.

The fourth layer of the architecture is a bi-directional recurrent layer. This layer includes two sets of hidden units: a set with backward recurrence $h^{(b)}$ and a set with forward recurrence, $h^{(f)}$:

$$\mathbf{h_t^{(f)} = g\ (W^{(4)}\ h_t^{(3)}\ t + W_r^{(f)}\ h\ _{t-1}\ ^{(f)} + b\ ^{(4)})} \quad (4)$$

$$\mathbf{h_t^{(b)} = g\ (W^{(4)}\ h_t^{(3)} + W_r^{(b)}\ h_{t+1}\ ^{(b)} + b^{(4)})} \quad (5)$$

Both the forward and backward units are passed as input to the fifth layer.

$$\mathbf{h_t^{(5)} = g(W^{(5)}\ h_t^{(4)} + b^{(5)})} \quad (6)$$

where, $h_t^{(4)} = h_t^{(f)} + h^{(b)}$

6. PERFORMANCE ANALYSIS

For the performance analysis of this model, we have used tensorboard functions and its visualizations.

6.1. Computing Average CTC loss and accuracy with different datasets

Epoch	Error Rate
0	0.953
10	0.747
20	0.794
30	0.761
40	0.725
50	0.694
60	0.641
70	0.593

Table.1. Decreasing Error rate with training steps.

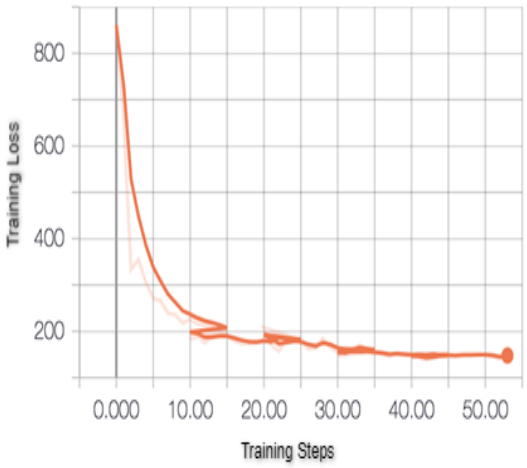

Figure.2. Training loss vs training steps

CTC can be understood as Connectionist Temporal Classification (CTC). We have used the CTC function

ISSN (Print): 2204-0595
ISSN(Online):22031731

available in the tensorflow library. It takes the logits and the target variables as inputs and computes the loss. Logits are basically the raw data that are obtained from the neural network's last layer. The average loss is minimized using the optimization functions like AdamOptimizer available in the tensorflow library.

From figure.2., the graph shows training loss vs training steps. It can be visualized that CTC is steadily decreasing with the training steps. Since the dataset, that we have used is not very large, it is recommended that to achieve good transcription accuracy (word error rate, CRT loss, and so on, which is speaker agnostic), we need to train it on a large dataset.

We have reduced the error rate to 0.593 in the final Epoch. Table 1. Depicts how the error rate reduced with training steps.

We have also tried to train the model with a different dataset. The alternate dataset that we took for the performance analysis of the model was the free-spoken digits audio dataset from GitHub. It is a simple audio/speech dataset that consists recordings of spoken digits in the format of .wav files at 8kHz. It consists of 1000+ files spoken by 6 speakers. The preprocessing and the feature extraction of the dataset is performed same as the previous dataset.
For this dataset, we achieved an accuracy of 0.9031 and we reduced the loss to 0.4391.

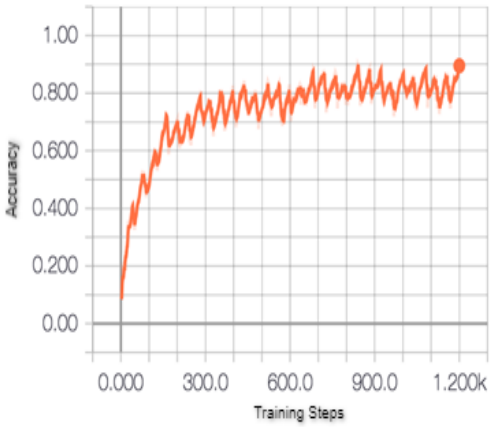

Figure.3. Accuracy vs Training steps for digit dataset

Figure.3. demonstrates that the accuracy of the model has successfully increased with the training steps.

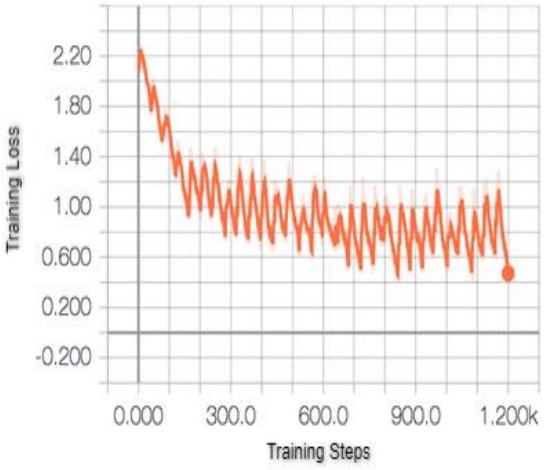

Figure.4. Loss vs Training steps for digit dataset

Figure.4. demonstrates that the loss of the model has successfully decreased with the training steps.

6.2. Comparison with other Techniques

Model	%WER
Vesely et al. [18] (GMM-HMM BMMI)	18.6
Vesely et al. [18] (DNN-HMM sMBR)	12.6
Maas et al. [19] (DNN-HMM SWB)	14.6
Maas et al. [19] (DNN-HMM FSH)	16.0
Seide et al. [20] (CD-DNN)	16.1
Kingsbury et al. [21] (DNN-HMM sMBR HF)	13.3
Sainath et al. [22] (CNN-HMM)	11.5
Soltau et al. [23] (MLP/CNN+I-Vector)	10.4
Proposed model (DeepSpeech)	**5.9**

Table.2. Published error rates (%WER).

Table.2. shows the %WER of different techniques. As illustrated in table.2., the proposed architecture gives a lower error rate than the above-mentioned other models and architectures. So, the proposed deepspeech

ISSN (Print): 2204-0595
ISSN(Online):22031731

model can be used for achieving lower word error rates.

7. CONCLUSION

Deepspeech architecture is proposed for the speech recognition systems. The proposed architecture is expected to give better results over other HMM based models. The main advantage of these neural network-based models over HMM based models is that neural network-based methods can learn end to end speech transcription entirely from data. The model works with better accuracy and lower word error rates than the HMM based models. The model is also expected to give better results than the traditional models on noisy data also. The only requirement of the proposed architecture is that it requires comparatively a larger dataset for the training purposes.

8. FUTURE ENHANCEMENTS

The future enhancements in our proposed architecture could be as follows:

a) The accuracy of the model can be improved by enhancing the architecture or training it on multiple GPUs with a larger dataset.
b) The model can also be trained to give better results for continuous speech recognition.

9. REFERENCES

[1] Krizhevsky, A., Ilya Sutskever and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks, In Proceedings of Communications of the ACM 60, pages 84-90, 2017.

[2] Ilya Sutskever, Oriol Vinyals and Quoc V. Le. Sequence to Sequence Learning with Neural Networks, In Proceedings of NIPS 2014, pages 3104–3112, 2014.

[3] Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals and Wojciech Zaremba. Addressing the rare word problem in neural machine translation, In Proceedings of ACL 2015, pages 11-19, 2015.

[4] K. Geetha and Dr. R. Vadivel. Phoneme Segmentation of Tamil Speech Signals Using Spectral Transition Measure, In Proceedings of Oriental journal of Computer Science and technology, Vol. 10, pages 114-119, 2017.

[5] Dhanashri, D. and S. Dhonde. Speech Recognition Using Neural Networks: A Review. In Proceedings of International Journal of Multidisciplinary Research and Development 2, pages 226-229, 2015.

[6] Veisi, Hadi & Mani, Armita. Persian speech recognition using deep learning, In Proceedings of International Journal of Speech Technology, pages 893-905, 2020.

[7] Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li and Yang Zhang. Quartznet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions, ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pages 6124-6128, 2020.

[8] Dinesh Kumar Vishwakarma, Ashima Yadav. Sentiment analysis using deep learning architectures: a review, In Proceedings of Artificial Intelligence Review, pages 4335-4385, 2019.

[9] T. Tanaka, R. Masumura, T. Moriya and Y. Aono, Neural Speech-to-Text Language Models for Rescoring Hypotheses of DNN-HMM Hybrid Automatic Speech Recognition Systems, In Proceedings of 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, pages 196-200, 2018.

[10] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno, and Tetsuya Ogata. Audio-visual speech recognition using deep learning, In Proceedings of Applied Intelligence 42, pages 722–737, 2015.

[11] Mohammad Mehdi Homayounpour, Toktam Zoughi, Mahmood Deypir, Adaptive windows multiple deep residual networks for speech recognition, In Proceedings of Expert System with Applications, Volume 139, 2019.

[12] Chengyi Wang, Yu Wu, Liang Lu, Shujie Liu, Jinyu Li, Guoli Ye, Ming Zhou. Low Latency End-to-

ISSN (Print): 2204-0595
ISSN(Online):22031731

End Streaming Speech Recognition with a Scout Network, In Proceedings of InterSpeech 2020 October 25–29, 2020, Shanghai, China, 2020.

[13] Soufane Hourri and Jamal Kharroubi. A deep learning approach for speaker recognition, In Proceedings of International Journal of Speech Technology, pages 123-131, 2020.

[14] Jinyu Li, Rui Zhao, Hu Hu and Yifan Gong. Improving RNN Transducer Modeling for End-to-End Speech Recognition, In Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), SG, Singapore, 2019, pages 114-121, 2019.

[15] Xiaoyu Qiu, Zhiquan Feng, Xiaohui Yang and Jinglan Tian. Multimodal Fusion of Speech and Gesture Recognition based on Deep Learning, In Proceedings of Journal of Physics: Conference Series, CISAI 2019, pages 6588-6596, 2019.

[16] Savitha G, Vibha L and Venugopal K R. Multimodal Biometric Authentication System Using LDR Based on Selective Small Reconstruction Error, Journal of Theoretical and Applied Information Technology, Vol.92. No.1 ISSN: 1992-8645, 2016.

[17] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567, 2014.

[18] Veselý, K, Ghoshal, A, Burget, L & Povey, D 2013. Sequence-discriminative training of deep neural networks, In Proceedings of Interspeech 2013. ISCA, 14th Annual Conference of the International Speech Communication Association, Lyon, 2013.

[19] A. L. Maas, A. Y. Hannun, C. T. Lengerich, P. Qi, D. Jurafsky, and A. Y. Ng. Increasing deep neural network acoustic model size for large vocabulary continuous speech recognition. abs/1406.7806, 2014.

[20] F. Seide, G. Li, X. Chen and D. Yu. Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription, In Proceedings of 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, pages 24-29, 2011.

[21] B. Kingsbury, T. N. Sainath, and H. Soltau. Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization, In Proceedings of Interspeech, 2012.

[22] Tara N Sainath, Abdel-rahman Mohamed, Brian Kingsbury and Bhuvana Ramabhadran. Deep convolutional neural networks for LVCSR, In Proceedings of 2013 IEEE international conference on acoustics, speech and signal processing, pages 8614-8618, 2013.

[23] H.Soltau, G.Saon and T.N.Sainath. Joint training of convolutional and non-convolutional neural networks,Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, pages 5572-5576,2014

ISSN (Print): 2204-0595
ISSN(Online):22031731