Turkish Journal of Computer and Mathematics Education Vol.12 No. 10 (2021), 7308-7315
Research Article

Recurrent network-based hybrid acoustic model for Automatic Speech Recognition

M.C.Shunmugapriya[1], D.Karthika Renuka[2], L.Ashok Kumar[3], J.Akila[4], K.Aneesha Banu[5], P.Priya Dharshini[6]
[1]Research Scholar, Department of IT, PSG College of Technology, Coimbatore, India.
[2]Associate Professor, Department of IT, PSG College of Technology, Coimbatore, India.
[3]Professor, Department of EEE, PSG College of Technology, Coimbatore, India.
[4,5,6]UG Student, Department of IT, PSG College of Technology, Coimbatore, India.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract
Speech is a key means of communication. Nowadays, speech is becoming a more common, if not standard, interface to technology. This will be seen within the trend of technology changes over the years. Increasingly, voice is employed to regulate programs, appliances and private devices within homes, cars, workplaces, and public spaces through smartphones and residentialassistant devices using Amazon's Alexa, Google's Assistant and Apple's Siri, and other proliferating technologies. This is often achievable with the help of Automatic Speech Recognition (ASR). Automatic Speech Recognition is a process that accurately translates spoken utterances into text. These technologies enable machines to reply correctly and reliably to human voices and supply useful and valuable services. As communicating with computer isquicker using voice instead of using keyboard, so people will prefer such system. Communication among the person is dominated by speech, therefore it's natural for people to expect voice interfaces with computer. This can be accomplished by developing speech to text which allows computer to translate voice request and dictation into text. The three models in traditional ASR system are acoustic model, language model and lexicon model. The challengesinvolved in Automatic Speech Recognition are different styles of speech, environment which include background noise and also accent of speaker. To mitigate these challenges, deep learning models are utilized. The main idea is to analyses features of input audio signals such as spectrogram and MFCC and to develop cutting edge deep learning models. The proposed end-to-end model achieved an error rate of 0.60 on Librispeech dataset.

Keywords: Automatic speech recognition, recurrent neural network, Deep learning, Word error rate

1 Introduction
Automatic Speech Recognition (ASR) aims to convert raw audio into sequence of corresponding utterance. An ASR system produces the foremost likely utterance sequence given a speech signal waveform. Speech is the most common form of communication. Automatic Speech Recognition is one among the important tasks within the Deep Learning field. Neural network forms the base of deep learning, a subfield of machine learning where the algorithms are inspired by the structure of a human brain. Deep neural network are such types of networks where the layer can perform complex operations suchas representation and abstraction that make the sense of sound, text, etc. The flexibility and predicting power of deep neural networks, which have recently become more accessible, are an advantage of deep learning for speech recognition.

A recurrent neural network that has been trained to ingest speech waveforms and produce English text transcriptions is the foundation of ASR. The aim of using recurrent neural network is to truly achieve an end-to-end deep learning system. The desirable building of Automatic Speech Recognition interfaces used both by literate and illiterate users. It is a very useful accessibility tool for people with hearing impairments. This speech recognition also preserves the endangered languages. Several sources of variability make ASR a difficult problem is environment (Background noise), speaker characteristics (Accent), etc. The solution is to fuse the acoustic model with a language model that's capable of understanding context. Ultimately, deep learning remains fairly in its infancy but is quickly approaching a state-of-the-art capability in speech recognition.

ASR system is composed of the following models

- ***Acoustic Model***: Converts raw speech to phonemes (the basic linguistic unit)
- ***Pronunciation Model***: Dictionary of phoneme sequence and their corresponding words
- ***Language model***: Probabilistic model which finds the most likely sentence.

2 Related Works

The Automatic Speech Recognition system for traffic control uses the Hidden Markov model (HMM) in feature extraction while its phraseology is predicated on the commands utilized in air applications [1]. They [2] have demonstrated that models trained with corpus do better on the standard Wall Street Journal (WSJ) test sets than models built on WSJ itself – the larger size of corpus (1000 hours, versus the 82 hours of WSJ's si-284 data) outweighs the audio mismatch. In this paper [3], they discussed the various techniques of Automatic Speech Recognition and Hidden Markov Model (HMM) of how the technology has progressed from the last years. They concluded that Compare ASR to the task of automatically driving a car; the latter requires intelligent interpretation of the field of vision for cameras mounted on a vehicle. While algorithms needed for cars would be very different for ASR, there are similarities in signal processing and both challenges seem daunting [4]. In this paper [5], they achieved good results frame accuracy with our CNN-classifier that leverages strong local correlation in speech signals.

Novel neural network architectures have been successful in both decoding words from phonemes and identifying phonemes from speech. In this paper, the author explained about Automatic Speech Recognition, the architecture of Deep Neural Network, Convolutional Neural Network and Recurrent Neural Network and their performances. [6]. Recurrent Neural Networks are considered to be the best algorithm for sequential data. It process and gives the predictive results due to its internal memory [7]. RNN is a network that has memory which decides the future predictions. It predicts one letter it will affect the likelihood of the upcoming letter which it will predict next one too. It uses the idea of sequential information. RNN, a neural network that has a memory that influences future predictions Sequential information which is stored in memory of RNNs is used for predictions [8] [9]. End-to-end training methods like Connectionist Temporal Classification (CTC) make it possible to train RNNs for sequence labelling problems where the input-output alignment is unknown. In [10][11][12] proposed deep recurrent neural networks, which combine the multiple levels of representation that have proved so effective in deep networks with the flexible use of long range context that empowers RNNs.

3 Methodology

3.1 Recurrent Neural Network

A Recurrent Neural Network (RNN) is a feed-forward neural network that is extended to a series of vector inputs. However, a memory of the previous time steps in the process must be retained in order to integrate temporal meaning into the next time step's forecast. In certain cases, understanding what will happen in future time steps will help guide the forecast at time t. Both the forward and backward contexts can be inserted into a forecast using bidirectional RNNs. This is achieved by running two RNNs in a sequence, one forward and one backward. The forward context RNN receives the inputs in forward order*t = {1,2,...,T}* for an input sequence *A = {a1,a2,...,aT},* and the backward context RNN receives the inputs in reverse order *t = {T,T-1,...,1}* for an input sequence *A = {a1,a2,...,aT}*. Together, these two RNNs form a single bidirectional layer. The two RNNs' outputs, h^f and h^r, are often combined to form a single output vector, either by summingthe two vectors, concatenating, averaging, or another way. The forward propagation for ht and the output prediction are given in equations 1 and 2.

$$ht = \tan h\,(Ua_t + Wh_{t-1}) \text{ ---------- (1)}$$

$$\hat{y}t = softmax(Vh_t) \text{---------(2)}$$

Where, the learnable parameters are U, W, and V. U incorporates the information from at,W incorporates the recurrent state, and V learns a transformation to the output size and classification. A diagram of this RNN is shown in Fig. 1.

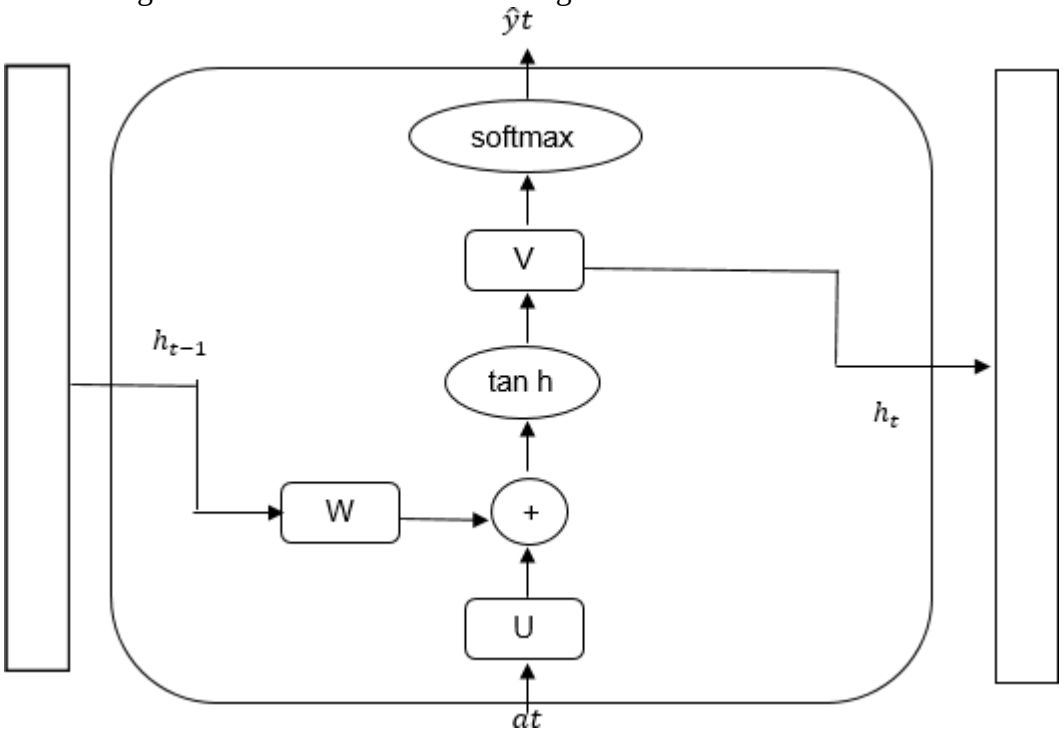

Fig. 1: Forward propagation of RNN

The gradients are calculated by weighing each direction that led to the ŷ forecast. Backpropagation over time is the term for this method (BPTT). The layers of the RNN-GRU is shown in Fig. 2.

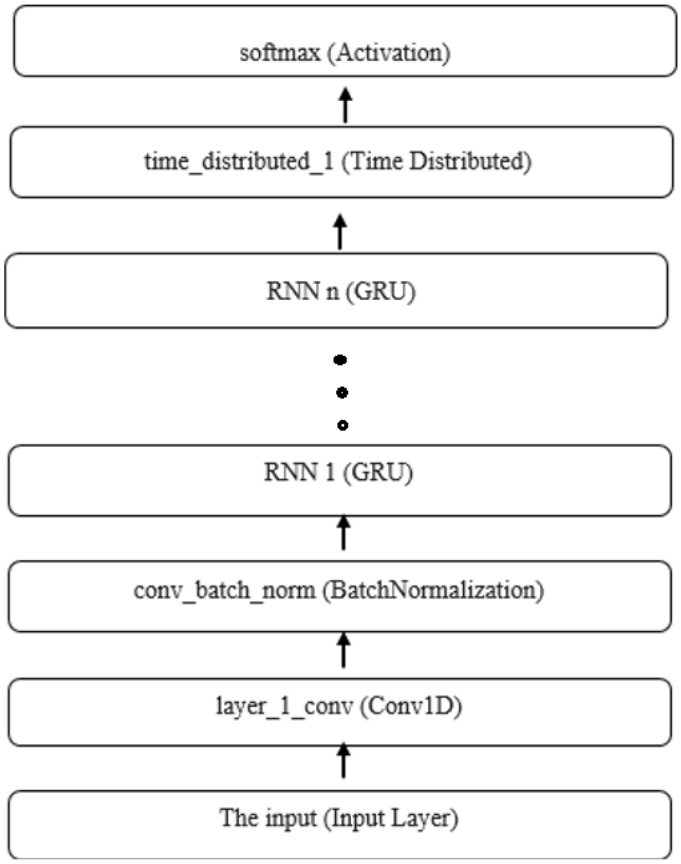

Fig. 2: RNN-GRU Layers

For RNN, the Gated Recurrent Unit (GRU) is a typical gating structure. The GRU integrates the gates in the LSTM to build a simplified upgrade rule with one less learned sheet, which decreases complexity and increases performance. The GRU blends the LSTM gates to build a simplified update rule with one less learned sheet, reducing complexity and increasing performance. The decision to use LSTM or GRU is primarily dependent on experience. The upgrade laws' equations are shown in equation 3 to 6.

$$Z_t = \sigma(W_z a_t + U_z h_{t-1}) \text{ ----------(3)}$$

$$r_t = \sigma(W_r a_t + U_r h_{t-1}) \text{---------- (4)}$$

$$\tilde{h}\text{t} = \tan h\,(W_h a_t + U_h h_{t-1}{}^{\circ} r_t) \text{ ---------- (5)}$$

$$h_t = (1 - Z_t)^{\circ}\, \tilde{h}\text{t} + Z_t * h_{t-1} \text{ ----------(6)}$$

4 Experimental analysis

The experiments are performed on LibriSpeech dataset. LibriSpeech Dataset consists of approximately 1000 hours of 16kHz read English speech, prepared by Vassil Panayotov with the assistance of Daniel Povey. The data is derived from read audiobooks from the LibriVox project. The training data contains around 960 hours of speech from read audio book recordings. The evaluation is on the clean dev and test sets, which each contains around 5.4 hours of speech.

The 2136 examples of training data is used to train an acoustic model, which is our base speech recognition system. Then use language model to enhance its performance.

Table 1: LibriSpeech Dataset

subset	hours	per-spk minutes	female speakers	male speakers	total speakers
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40

The evaluation is on the clean dev and test sets, which each contains around 5.4 hours of speech. The first step is that the pre-processing step that converts raw audio to atleast oneof two feature representations that are commonly used for ASR - MFCC, Spectrogram as shown in fig 3 and fig 4. Theidea behind MFCC features is: the recorded speech signals are sampled and stored using Audacity. The sampling is completed at a rate of 16000 samples per second. Each speech signal is split into windows of 16 ms each and hence, 256 samples each. A spectrogram is a visual way of representing the signal strength, or "loudness", of a signal over time at various frequencies present during a specific waveform. Not only canone see whether there's more or less energy at, for example, 2 Hz vs 10 Hz, but one also can see how energy levels vary over time.

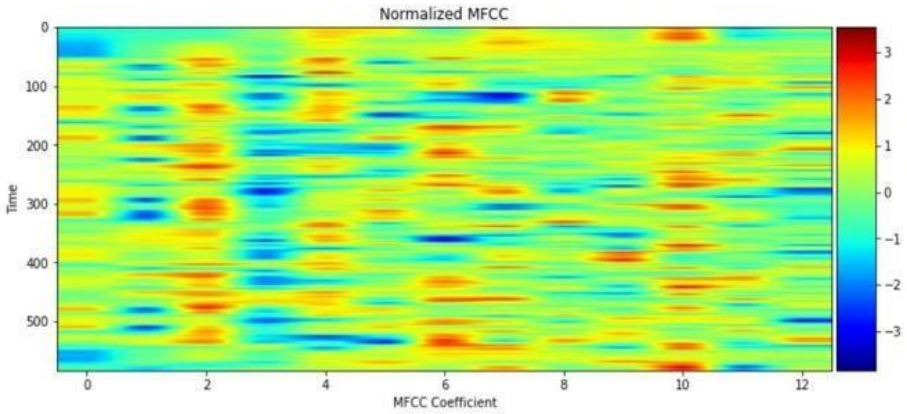

Figure 3: Mel Frequency Cepstral Coefficients (MFCCs)

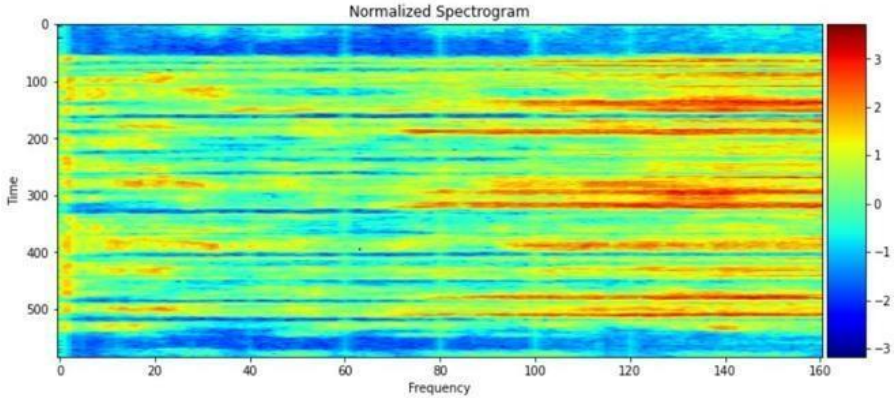

Figure 4: Spectrogram

Acoustic model which accepts audio features as input and returns a probability distribution over all potential transcriptions RNN-GRU model with the CTC loss criterion is used for train the model. CTC is just a loss function that is usedto train Neural Networks, like Cross-

Entropy and so on. It's used at problems, where having aligned data is a problem, like Speech Recognition. Forty Mel-scale filter bank coefficients and their delta and delta-delta features are concatenated as their input features.

The baseline RNN model is a five-layer GRU with 250 cells in each layer and direction. Dropout rate of 0.2 is applied on each layer with Relu activation. Batch Normalization is added to the recurrent layer to scale back training times and it is also be used to standardize inputs before or after the activation function of the previous layer. The Time Distributed layer is employed to seek out more complex patterns within the dataset. The model is trained over 30 epochs with SGD as optimizer. The training loss after 30 epochs is 64.7593 and the validation loss is 110.0321. The validation loss starts decreasing then it began to increase. This means that the network had started to over fit to the training set.

4.1 Performance Metrics

The performance metric is the training and validation accuracy loss. The purpose of loss functions is to compute the quantity that a model should seek to minimize during training. WER is used to measure the performance of the speech recognition.

4.1.1 Visualization Graph

The training and validation loss is plotted over 30 epochs for the Neural Networks. As depicted in below figure 5, it is evident, that the train and test loss increases and decreases over epochs.

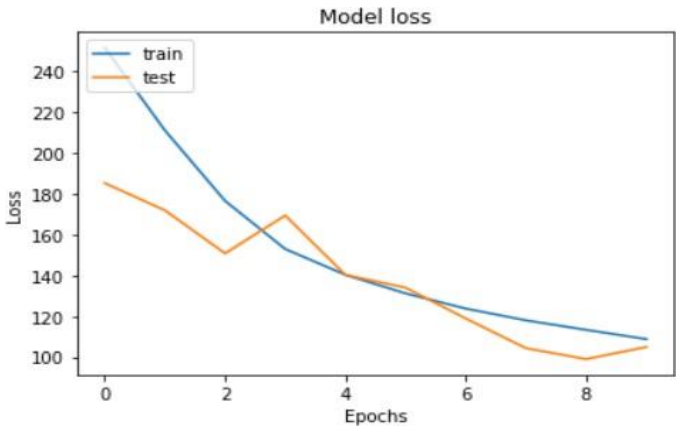

Figure 5: Model Loss Plot

4.1.2 Word Error Rate (WER)

Word error rate (WER) could be a common metric of the performance of a speech recognition or artificial intelligence system.
The general problem of measuring performance lies within the indisputable fact that the recognized word sequence will have a distinct length from the reference word sequence. The WER comes from the Levenshtein distance, engaging at the word level rather than the sound level. The WER could be a valuable tool for comparison completely different systems also as for evaluating enhancements among one system. Word error rate can then be computed as in equation

$$WER = \frac{S+D+I}{N} \text{ ---------- (1)}$$

Where, S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, N is the number of words in the reference (N=S+D+C)

Our model is compared with baseline models as given in table 2. When trained end-to-end with suitable regularization, RNN-GRU model achieved an error rate of 0.60 on the Librispeech dataset.

Table 2: Word Error Rate Comparison with baseline model

Models	**WER**
Jasper DR 10x5 (with LM)	0.29
RNN-GRU	0.60

5 Conclusion

Automatic Speech Recognition is the challenging problem to deal with. This work explores advancements in recent literature regarding the replacement of the GMM- HMM based automatic speech recognition system with a deep neural networks. Speech recognition has created a technological impact on society and is expected to flourish further inthis area of human machine interaction. An ASR system uses acoustic models to extract information from the acoustic signal. Hence, acoustic models basically build using the probability distribution over the acoustic space. Nowadays end to end speech recognition systems are more common which implement RNN as their fundamental algorithm. It can be seen that end-to-end speech recognition greatly simplifies the complexity of traditional speech recognition. RNN GRU model had achieved 0.60 WER. In future, Language model willbe implemented to enhance the performance of ASR (i.e reduce the WER).

Acknowledgement

Our sincere thanks to Department of Science and Technology, Government of India for funding this project under - Department of Science and Technology Interdisciplinary Cyber Physical Systems (DST – ICPS) scheme.

References

[1] Pratiksha.C.Raut, Seema.U.Deoghare, "Automatic Speech Recognition and its Applications", International Journal of Engineering and Advanced Technology, vol. 3, no.5, May 2016

[2] Vassil Panayotov, Guoguo Chen, Daniel Povey, Sanjeev Khudanpur, "Librispeech:
An ASR corpus based on public domain audio books",2019
[3] Anchal Katyal, Amanpreet Kaur, Jasmeen Gill, "Automatic Speech Recognition: A Review," International Journal of Engineering and Advanced Technology, vol. 3, no. 4, july 2015
[4] Akhilesh Halageri, Amrita Bidappa, Arjun C, Madan Mukund Sarathy, Shabana Sultana, "Speech Recognition using Deep Learning", International Journal of Computingand Knowledge, Vol. 6, no. 3, 2015
[5] Song, W. (2015). End-to-End Deep Neural Network for Automatic Speech Recognition.
[6] Lekshmi.K.R, Dr.Elizabeth Sherly, "Automatic Speech Recognition using different Neural Network Architectures – A Survey", International Journal of Computing and Knowledge, Vol. 7, no. 6 , 2016
[7] Sruthi Vandhana, Srivibhushanaa S, Sidharth K, Sanoj C S, "Automatic Speech Recognition using Recurrent Neural Network", International Journal of Engineering Research & Technology (IJERT), Vol. 9, Issue 08, August-2020
[8] Aditya Amberkar, Parikshit Awasarmol, Gaurav Deshmukh, Piyush Dave, "Speech Recognition using Recurrent Neural Networks", 2018
[9] Dr.R.L.K.Venkateswarlu, Dr. R. Vasantha Kumari, G.Vani JayaSri , "Speech Recognition By Using Recurrent Neural Networks", International Journal of Scientific &Engineering Research Volume 2, Issue 6, June-2011
[10] Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton, "Speech Recognition With Deep Recurrent Neural Networks", 2013
[11] Dr. S Lovelyn Rose, Dr. L Ashok Kumar, Dr. D Karthika Renuka, "Deep Learning using python", Wiley India.
[12] Dr. S Lovelyn Rose, Dr. L Ashok Kumar, Dr. D Karthika Renuka, "Design and Implementation Of E-Learning System Using Deep Learning Based On Audio-Video Speech Recognition For Hearing Impaired", Perspectivas em Ciencia da Informacao, Vol 22,pp 192, 2017.