Automatic Speech Recognition for Indian Accent Lectures contents using End-to-End Speech Recognition model

Dr.L.Ashok Kumar[1], Dr.D.Karthika Renuka[2], G.Raajkumar[3]
{lak.eee@psgtech.ac.in[1],dkr.it@psgtech.ac.in[2], graajkumar97@gmail.com[3]}

Professor, Department of Electrical and Electronics Engineering, PSG College of Technology[1],Associate Professor, Department of Information Technology, PSG College of Technology[2],Junior Research Fellow, Information Technology, PSG College of Technology[3]

Abstract. In a variety of voice search applications, Automatic speech recognition (ASR) systems are used. The process of turning speech to text is known as automatic speech recognition (ASR). Most of the ASR research is happening using American and British accent. Hence, in this work we have made an attempt to convert Indian accent speech to text using NPTEL lecture audio. The proposed work involves speech to text using deep learning models for Indian accent speech. LAS has two main components one is based on sequence-to-sequence framework with a pyramid structure, by reducing the encoder steps in number the decoder must attend through the end-to-end process. The result obtained from the proposed work improve Word Error Rate of 14%.

Keywords: Automatic speech recognition (ASR), Indian accent, Word Error Rate (WER), NPTEL lecture audio, Listen, Attend, and Spell (LAS).

1 Introduction

The Listen, Attend, and Spell (LAS) model is a well-known architecture for proceeding with end-to-end automatic speech recognition (ASR). LAS results in producing high accuracy than existing system proposed so far. It has beaten the deep neural network based Markov model by solving large scale ASR problems. The challenge also imitates some problem grown during the output creation and in input synchronization in which it won't support well so the problem of online recognition still remains a problem. In recent years, analysis has progressed with alternatives like the RNN electrical device (RNNT), which permits for time-synchronous coding, multi-model techniques, within which a time-synchronous model like RNN-T or bureau provides temporal arrangement data to the LAS decoder or specially designed loss functions that punish latency. These techniques are effective, though they will necessitate a lot of difficult coaching algorithms or coding procedures.

In this paper, a neural network model called (LAS) is employed, improve on previous attempts [10]. The trained neural network to convert an audio stream into a word sequence character by character. Unlike previous approaches, LAS does not rely on HMMs or make assumptions about label sequence independence. LAS is attention-based learning. Both listener and a speller recurrent neural network (RNN). Because they take in low-level voice inputs and translate them into higher-level features, pyramidal RNNs are utilised in the listener [10]. The speller is a recurrent neural network (RNN) that employs the attention

ICCAP 2021, December 07-08, Chennai, India

DOI 10.4108/eai.7-12-2021.2314531

mechanism to define a probability distribution over character sequences and then converts that probability distribution into output utterances. While using RNN pyramidal listener model it will reduces the no of step in attention model. This is really important to our overall approach. Rare words and vocabulary (OOV) words are automatically execute in the model. Because the character sequence will be produce only one character at a time. Another advantage in employing character model, the character may generate variety of spelling variants [11].

The contributions of this paper includes:

1) Use of NPTEL Indian accent "dataset" to minimise error rate.
2) In order to improve the accuracy of collecting suitable words for the lecture capture system, the LAS model will be employed.

2 Related Work

Deep Neural Networks (DNNs) have developed various aspects of acoustic recognition software. They're often used DNN-HMM recognition systems for audio modeling [2, 3]. Pronunciation models, which map words to phoneme sequences, have also improved significantly as a result of DNNs [4, 5]. Enhance voice recognition by using n-best under recurrent model will be improved value of lists [6]. Audio, pronunciation, grammar and language are the four models are taught separately. In recent research the problem of disjointed training by constructing models that are trained from speech to transcripts end-to-end [7, 8]. Both Connectionist Temporal Classification (CTC) and sequence to sequence models approaches. However the sequence to sequence technique has only been applied to phoneme sequences.

Deep neural networks solves many complex applications in that case it involves in high classification rate i.e., conversion of fixed-length vector to an output classes [12].Structured problem were increased in text classification to address the problem mapping the variable length to another a term called CRFs, it will be revolves around Hidden Markov Models (HMMs) [13] with conditional Random Fields. The Models results in end-to-end length and makes native assumptions about the probability of data.

Sequence to sequence is a framework that tries to solve the problem of learning the input and output sequences, for converting the variable length input into fixed length vector RNN encoding is used. The designed model used beam search for interfacing the candidates for next step predictions. Model can be improved by adding the attention mechanism that sends additional information about the RNN decoder to generate the output in token form. In output state the layer decodes the RNNs previous hidden state to build an attention vector for spanning the sequence of input [17]. Transferring a data from decoder to encoder at a same time is done by attention vector. To increase the flow and to pass the information securely attention vector that skip the link through the RNN for not losing the loosing with holding effectively.

3 Methodology

Acoustics features are fed into Listen, Attend, and Spell model as input sequence and provides outputs in the form of English characters. Let x = (x_1, . . . , x_T) be our filter bank spectral feature input sequence, and let y = (<sos>, y_1, . . . , y_s,<eos>), yi∈ {a,b,c,⋯,z,0,⋯,

9,<space>,<comma>,<period>,<apostrophe>,<unknown>}, be the characters in output sequence. The unique start-of-sentence token is represented as <sos> and end-of-sentence token is represented as <eos>.

Using the chain rule, models every character output as y_i in a conditional distribution for the previous character$y_{<i}$, x as an input signal.

$$p\left(\frac{x}{y}\right) = \Pi p(\frac{y_i}{x}, y_{<i}) \quad (1)$$

There are two sub-modules in the LAS model: The one sub module is the listener and another one is the spell. An acoustic model encoder serves as the listener. It is responsible for the crucial task of listening. Attention-based character decoder serves as the speller. It is responsible for the task of attend and spell. The Listener function transforms the input signal x into a high level signal. The high level signal's representation as h = $(h_1, \ldots, h_U)$ with U ≤ T, The Attend And Spell function, on the other hand, uses h to construct a probability distribution across character sequences:

$$\mathrm{h} = \mathrm{Listen}\ (\mathrm{x}) \quad (2)$$

$$\mathrm{p}\left(\frac{\mathrm{y}}{\mathrm{x}}\right) = \mathrm{Attend\ and\ Spell}\ (\mathrm{h}, \mathrm{y}) \quad (3)$$

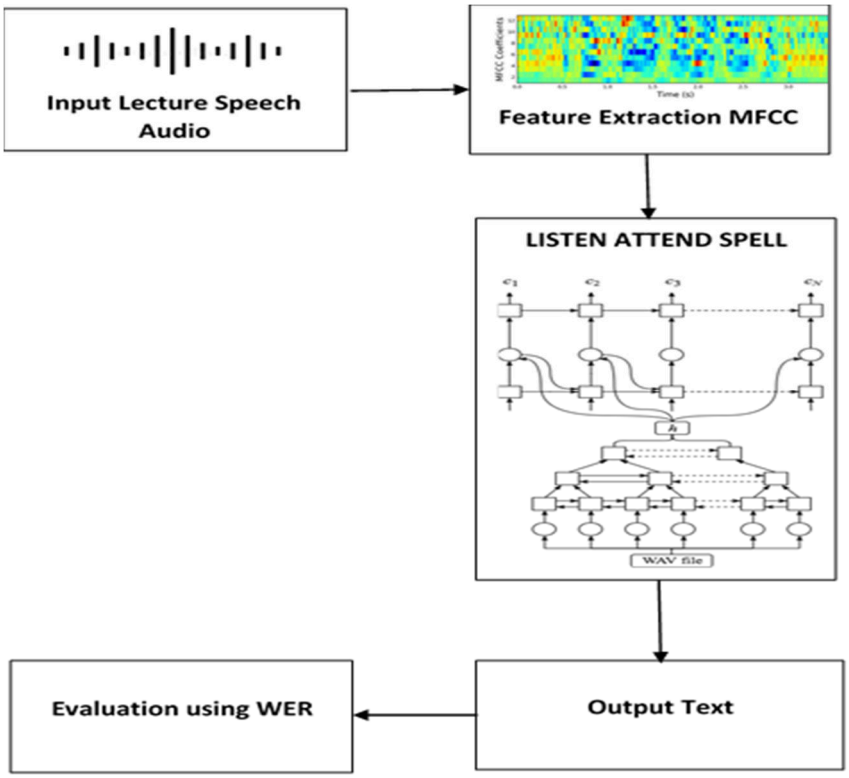

Figure 1. Proposed model for speech recognition

The input sequence x is encoded into high-level options h by the observer, and also the speller produces y characters from h.

A RNN with a pyramid is utilised for bidirectional Long Short term memory. Since the length of the input audio signal can range from 100 to 1000 of frames, such modification is essential

to reduce the length U of h from T. After a month of training, one simple implementation of BLSTM for the activity Listening resolved gradually and produced results that have been worse to all those described below. This is most likely because Attend and Spell has trouble retrieving relevant data from a high set of input clock cycles. Decrease the temporal resolution by a factor of 2 in each subsequent stacked PBLSTM layer. The output from the j-the layer in BTLM architecture is computed as follows at the

$$h_i^j = PBLSTM\left(h_{2i}^{j-1}, h_{2i+1}^{j-1}\right) \quad (4)$$

For reducing the ration from 2:3 into 8 times the model suggested to implement a three PBLSTMs developed on the top of BLSTM layer. As a result, the attention model is able to extract meaningful data from fewer time steps. The deep design allows the model to learn nonlinear feature data representations, in addition to reduce resolution. An illustration of the PBLSTM is shown in Figure 1. Its convergence speed is also reduced by the hierarchical order.

LSTM Transduce is use in attention based system, in each output state the transduce that generates a probability distribution across all the characters based on the pervious works. For decoding s_i and the context c_i used to determine the movement of y_i

Decoder state s_i is implies on proceedingc_{i-1}, the previously used character y_{i-1}, and the context c_{i-1}., i in the equation states an attention mechanism.

$c_i = Attention\ Context\ (s_i, h)$ (5)

$\mathrm{s_i} = \mathrm{RNN}\ (\mathrm{s_i} - 1, y_i - 1, c_i - 1)$ (6)

$\mathrm{p(y_i/x, y_{<i}) = Character\ Distribution(s_i, c_i)}$ (7)

Character Distribution is a Multilayer perception with softmax as an activation function, output layer for characters with RNN in two-layer LSTM. Attention Context generate vector with content, c_i encapsulating the information to generate next character in acoustic signal.

Scalar energy implies the attention function of $\mathrm{e_{i,u}}$, it contains each decoder time step of u of vector $\mathrm{h_u} \epsilon \mathrm{h}$ and$\mathrm{s_i}$.Activation function of softmax with scalar energy $\mathrm{e_{i,u}}$is is converted into the probability distribution over each time steps of attentionα_i.

$e_{i,u} =< \emptyset(s_i), \varphi(h_u) >$ (8)

$\alpha_{i,u} = \frac{\exp(e_{i,u})}{\sum_u exp(e_{i,u})}$ (9)

$\mathrm{c_i} = \sum_\mathrm{u} \alpha_\mathrm{i,u} \mathrm{h_u}$ (10)

Φ and ψ are MLP networks, α_i distribution in sharp, h for frames; continuous bag of weighted features of h is forc_i. LAS architecture is explained in Figure.1.

Proposed work contains 3 layers of 512 PBLSTM with 256 nodes per direction on BLSTM top present on Listen function for input layer. Reducing the time resolution from 8 to 23 for the result. LSTM with two layers make use of spell function with 512 nodes for each layer. The proposed model was trained using Asynchronous Stochastic Gradient Descent (ASGD). A learning rate of 0:2 and a geometric decay of 0:98 per 3M utterances (i.e., 1=20-Th of an epoch) were used. The DistBelief framework was used with 32 replicas, each with a mini batch of 32 utterances. To expedite training, sequences were divided into buckets depending upon their frame length.

4 Result Analysis

Dataset * collection involves collecting 1000 lecture audio files from NPTEL as a source hub. The dataset were separated into 80:20 ratio as a count 750 speech data were added to training set and remaining 250 speech data to test all the existing files were in wav format. A sample of lecture speech is listed in Table 1 as an illustration.

*- Link to reach out the dataset https://github.com/AI4Bharat/NPTEL2020-Indian-English-Speech-Dataset

Table 1. Sample lecture content text form NPTEL

S.No	Lecture Content
1	This entire flow towards the other direction until and unless there is boundary layer.
2	Such that y x satisfies this equation
3	Especially in the explosive environment this is preferred.

5 Evaluation Metric

The word error rate used for measuring the performance of speech recognition system as defined in equation 11,

$$WER = \frac{S+D+I}{N} (11)$$

Where, S = Total substitution

D = Total deletions

I = Insertion count

N = Total words in reference

Table 2. The perform of LAS model on different datasets

Dataset	Accent	WER
Librispeech	American	2.8%
NPTEL	Indian	14%

Table 2 shows the word error rate of the proposed system using NPTEL audio files with comparison among the other existing methods.

6 Conclusion

The proposed system includes implementation of Listen, Attend, and Spell (LAS) on NPTEL Indian accent speech achieved an improved performance. LAS has two main components one is based on sequence-to-sequence framework with a pyramid structure, by reducing the encoder steps in number the decoder must attend through the end-to-end process. Our experiment result shows that the WER of 14% is obtained using Indian lecture speech data on LAS model. Future work will be focus on increasing the number of files to train and

implement. As far now there is a small amount of files were taken to implement. As a result, future research will concentrate on massive datasets in order to improve accuracy.

References

[1] India, M., Safari, P. and Hernando, J., 2019. Self-multi-head attention for speaker recognition. arXiv preprint arXiv:1906.09890.

[2] NavdeepJaitly, Patrick Nguyen, Andrew W. Senior, and Vincent Vanhoucke. Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition. In INTERSPEECH, 2012.

[3] Tara Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and BhuvanaRamabhadran. Deep Convolutional Neural Networks for LVCSR. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[4] Kanishka Rao, Fuchun Peng, HasimSak, and Francoise Beaufays. Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2015.

[5] Kaisheng Yao and Geoffrey Zweig. Sequence-to-Sequence Neural Net Models for Graphemeto-Phoneme Conversion. 2015.

[6] Chan, W., Jaitly, N., Le, Q. and Vinyals, O., 2016, March. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4960-4964). IEEE.

[7] Hsiao, R., Can, D., Ng, T., Travadi, R. and Ghoshal, A., 2020. Online automatic speech recognition with listen, attend and spell model. IEEE Signal Processing Letters, 27, pp.1889-1893.

[8] L. Li, Z. Tang, D. Wang, and T. F. Zheng, “Full-info training for deep speaker feature learning,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5369–5373.

[9] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey Hinton. Acoustic modelling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing, 20(1):14–22, 2012.

[10] Alex Graves and NavdeepJaitly. Towards End-to-End Speech Recognition with Recurrent Neural Networks. In International Conference on Machine Learning, 2014.

[11] Prabha, S. Lavanya, M. Surendar, and M. Neelamegam. "Experimental investigation of eco-friendly mortar using industrial wastes." Journal of Green Engineering 9.4 (2019): 626-637

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In Neural Information Processing Systems, 2012.

[13] Ilya Sutskever, OriolVinyals, and Quoc Le. Sequence to Sequence Learning with Neural Networks. In Neural Information Processing Systems, 2014.

[14] C. Amuthadevi, D. S. Vijayan, Varatharajan Ramachandran, “Development of air quality monitoring (AQM) models using different machine learning approaches”, Journal of Ambient Intelligence and Humanized Computing, https://doi.org/10.1007/s12652-020-02724-2

[15] Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, OriolVinyals, and WojciechZaremba. Addressing the Rare Word Problem in Neural Machine Translation. In Association for Computational Linguistics, 2015.

[16] Sebastien Jean, Kyunghyun Cho, Roland Memisevic, and YoshuaBengio. On Using Very Large Target Vocabulary for Neural Machine Translation. In Association for Computational Linguistics, 2015.

[17] W. Cai, J. Chen, and M. Li, “Exploring the encoding layer and loss function in end-to-end speaker and language recognition system,” in Proc. Odyssey 2018 The Speaker and Language Recognition Workshop, 2018, pp. 74–81.