The 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)
Taipei, Taiwan, September 24–26, 2020. The Association for Computational Linguistics and Chinese Language Processing

Nepali Speech Recognition Using CNN, GRU and CTC

Bharat Bhatta
brb4344@gmail.com

Basanta Joshi
basanta@ioe.edu.np

Ram Krishna Maharjhan
rkmahajn@ioe.edu.np

Department of Electronics and Computer Engineering
Pulchowk Campus, Institute of Engineering
Tribhuvan University
Nepal

Abstract

Communication is an important part of life. To use communication technology efficiently we need to know how to use them or how to instruct these devices to perform tasks. Automatic speech recognition plays an important role in interaction with the technology. Nepali speech recognition involves in conversion of Nepali speech to its correct Nepali transcriptions. The purposed model consists of CNN, GRU and CTC network. The feature in the raw audio is extracted by using MFCC algorithm. CNN is for learning high level features. GRU is responsible for constructing the acoustic model. CTC is responsible for decoding. The dataset consists of 18 female speakers. It is provided by Open Speech and Language Resources. The build model can predict the with the WER of 11%.

Keywords: Nepali Speech Recognition, Automatic Speech Recognition, Gated Recurrent Unit (GRU), Convolution Neural Network (CNN)

1 Introduction

Speaking and writing are the two important things that help us to communicate among us. Deficient of either writing or speaking affects our daily activates. Most of the people in rular area are able to speak properly but not able to write properly. Most of communication technology (gadgets, mobiles, computers etc) needs text as an input for their operation. To make familiar with the technology Automatic Speech Recognition (ASR) can play significant role. The Nepali ASR converts the spoken Nepali voice to its textual representation.

The ASR can built by two different approaches. The first approach is traditional based that implement Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM). The input feature vector is efficiently processed by the GMM and calculated the emission probabilities for each HMM states[1]. The second is to deep learning approach to build the acoustic model. The use of deep learning approach significantly increases the performance of the ASR system[2]. Traditional ASR system needs separate block of phonetic/linguistic constructs, acoustic model and the language model. The requirement of separate phonetic/linguistic constructs is eliminated by the deep learning approach[3].

Previously, a work is carried out to recognised the ten Nepali unique words[4] and another work is carried out taking the dataset of sports news. It consists of 1320 words among which 617 are unique[5]. The second model could not recognised the characters 'दु', 'वि', 'के', 'लि' that occurs very close together suggesting that the network learned them as single sounds although they are multiple characters. The CNN is used to learn the features that can easily distinguish those close words.

This paper presents an idea to build the Nepali ASR system that can convert spoken Nepali language to its textual representation. The model used MFCC as input feature vector. These MFCC features area used by CNN to generate more spatial features. CNNs are used beacuse they are exceptionally good at capturing high level features in spatial domain[6]. GRU is used to develop an acoustic model. Training duration for GRU is less compared to LSTM network. CTC is used because it is alignment free[3]. The loss function used is CTC loss and decoding is carried out thorough the CTC network. The use of CTC eliminate the use of framewise labeling.

2 Review

Starting from single speaker based digit recognizer the modern Automatic Speech Recognition (ASR) reaches to speaker independent Hidden Markov Model based ASR[7]. With evolve of the deep learning the accuracy of the ASR system further increases[8]. Deep Neural

Network (DNN) domination in ASR started, which showed that feed-forward DNN outperforms (Gaussian Mixture Model) GMM in the task of estimation of context-dependent HMM state emitting probabilities[9].

The development of ASR for speech recognition passes through series of steps. Development of ASR starts from digit recognizer for single user , passing through HMM, GMM based and reaches to deep learning[10, 9]. Some research work has been carried on Nepali speech recognition and Nepali speech synthesis. The initial work on Nepali ASR is carried out by using HMM based approach. This system is trained with 10 different words. Four female and four male speakers record the data. This models is able to predict limited words only[4]. Since limited word based ASR system is not able to generalised unseen words. Increasing the size of vocabulary and building own dataset a model is built. The model is not able to predict the some words 'डु', 'वि', 'के', 'लि' that occurs very close together suggesting that the network learned them as single sounds although they are multiple characters[5]. To eliminate the wrong prediction on close character a CNN layer is added. The accuracy of the Nepali ASR model can be further increased by using n-gram language model. The n-gram model, which defined the probability of occurrence of an ordered sequence of n words[11].

3 Method

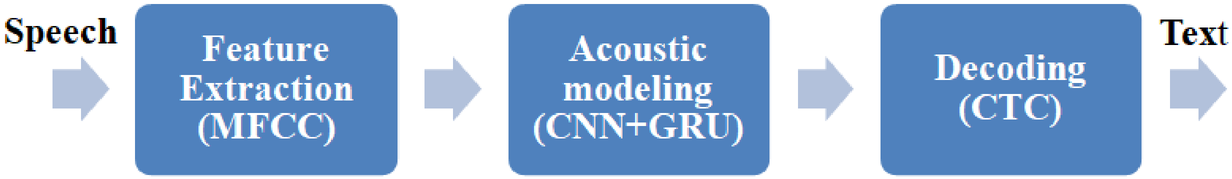

Figure 1: Architecture of proposed ASR system

The proposed ASR system involves in the conversion of speech input feature vector $X = [x_1, x_2, x_3, , , x_n]$ into its textual representation (label)$Y = [y_1, y_2, y_3, , , y_m]$ as shown in in Figure 1. The label is group of Nepali characters. The MFCC feature vector x_t of dimension D at time frame corresponds to label sequences $L = [_1, L-2, L_3, , , L_N]$. If vocabulary of labels is V, $l_u \epsilon V$ is the label at position u in L, then V^* represents the collection of all label sequences formed by labels in V. From these information the ASR need to find most likely label sequence $\hat{L}$ for given X This can be represented by the equation 1:

$$\hat{L} = \frac{argmax}{L \epsilon V^*} \; p(L/X) \tag{1}$$

The the main objective of an ASR is to established a model that can accurately calculate the posterior probability p(L/X).

3.1 Feature Extraction

Features are the elements that represents the phonemes in the speech. The features presents in the raw audio speech is is extracted by using the MFCC feature extraction algorithm. MFCC features are sequence of Acoustic feature vectors where each vector representing information in a small time window of signal[12]. The first step is pre-emphasis of the signal. The emphasis is carried out by choosing the value of $\alpha = 0.97$. Pre-emphasis boost the amount of energy present in high frequency signal. A signal is said to be stationary if its frequency or spectral contents are not changing with respect to time. Speech signal are non-stationary in nature because its spectral is constantly changing. So to simplify things it is assumed that on short time scales the audio signal doesn't change much, i.e. statistically stationary, obviously the samples are constantly changing on even short time scales[13]. Windows is taken to make them stationary signal[14]. The size of the window is 25 ms. 512 point FFT is taken from windowed signal. Thus obtained signal now model with human auditory system.26 filter-bank is implemented as the set of triangular shaped band-pass filters arranged in non-uniform frequency scale. This MFCC filter-bank is responsible to make them similar to human auditory system[15] and is converted to Log scale because human are less sensitive to higher energy change.

3.2 Feature Learning

CNN is used to capture high level spatial features from the image. The plot of MFCC can be view as a transformed intensity of frequencies over time which resembles to images[6], hence CNN can be used to capture high level feature in spatial domain. 1-dimensional CNN of filter size 200, kernel size (K) 11 and dilation 1 is used. Compared with conventional speech features, CNN can use local filtering and maximum pooling techniques to get more robust features[16].

3.3 Acoustic model generation

Gated Recurrent Unit (GRU) is used to learn the sequential data[17]. GRU manage the input weights to solve the vanishing gradient that present in RNN and consists of two gates: reset gate and update gate[18] which can be represented by the Egn. 2 and 3 respectively.

$$z_t = \sigma(W_z[h_t - 1, x_t]) \tag{2}$$

$$r_t = \sigma(W_r[h_t - 1, x_t]) \tag{3}$$

The reset gate is responsible for determining how much amount of past information to be forgotten. The update gate is responsible for determining how much amount of past information should be forwarded to for future[19]. The current memory and the final memory at current time step is given by the Egn. 4 and 5 respectively.

$$\hat{h} = \tanh(r_t.W_r[h_t - 1, x_t]) \tag{4}$$

$$h = (1 - z_t) * h_t - 1 + z_t * \hat{h_t} \tag{5}$$

3.4 CTC layer

The decoding is carried out is using CTC network. The CTC is based on the Bayes' on decision theory[2]. It receives output from softmax function. For each output layer, posterior probability (that represents 89 symbols) is computed. The character (symbol) having highest probability is given as the output. The decoding is carried out by the CTC network[20]. The output is generated in the form of numeric form. This output is mapped to the character. This is predicated output. CTC loss is computed by using true label and predicated output.

3.5 Evaluation

The Word Error Rate (WER) and Character Error Rate (CER) indicate the amount of text that the applied model did not read correctly. WER is the common evaluation metric for speech recognition system[21] which lies in the range between 0% and 100%.

4 Experiment

The experimental setup is carried out on the GPU MX150. For the pre-processing, feature extraction, training and testing, python and its library has been used.

4.1 Dataset

Speech recognition need the label data. It consist of audio corpus and its label file[8]. Every voice clip consists of phoneme transcription aligned with the sentence transcription label. The dataset consists of high quality female spoken corpus which is provided by Open Speech and Language Resources[22]. The dataset contains Eighteen unique female speaker records. The dataset is divided as 80% train and 20%. test set.

4.2 Result and Analysis

At first MFCC feature is taken from the raw audio. These features are passed to CNN to extract high level features. These features are used to generate the acoustic model. Several experiments has been conducted by varying several hyperparametser such as batchsize, learning rate, number of epochs. The size of CNN filter is 200, kernal size is 11 and in GRU units is 200.

The first experiment is carried out keeping 16412 as training utterance out of 26000 utterances. The learning rate is 0.03, momentum is 0.9, batch size is 100 and total number of epochs is 400. Total training duration is 1.5 days. The training is carried out in CPU Intel Core i7-8550U. The model get overfitted. It can predict well for train data but doesn't predict well for unseen data.

The second experiment is carried out keeping 16412 as training utterance out of 2064 utterances. The learning rate is 0.03, momentum is 0.9, batch size is 300 and total number of epochs is 100. Total training duration is 1.5 days. The other conditions remain unchanged.

The third experiment is carried out keeping 16412 as training utterance out of 2064 utterances. The learning rate is 0.015, momentum is 0.9, batch size is 50 and total number of epochs is 100. Total training duration is 1.5 hrs. The other conditions remain unchanged. The Batch size is changes to observe the effect. The result can be summarised by the Table 1. The parameters from this experiment is considered and some sample outputs are tabulated as shown in Table 2.

Table 1: Summary of experiments with the results

Experiment	learning rate	batch size	total epochs	WER
1	0.03	100	44	90
2	0.03	300	100	80
3	0.015	50	100	11

4.3 Model validation

The performance of the model is validate with the RNN-CTC model [5]. This research work is carried out to enhance the performance of the existing model. The Character Error Rate (CER) of that model is 52% and our CNN-GRU-CTC model gives the CER of 1.836%. Based on these results, it can be concluded that our model provides the better generalization capabilities.

Table 2: Sample output of the System

Sample	Ground Truth	Model prediction	WER
1	सानैदेखि सङ्गीतमा रुचि राख्ने अधिकारीले सुरुमा हार्मोनियममा शिव शङ्करबाट तालिम लिएका थिए	सानैदेखि सङ्गीतमा रुचि राख्ने अधिकारीले सुरुमा हार्मोनियममा शिव शङ्करबाट तालिम लिएका थिए	0.00
2	इन्डोनेसियाली पपुवा प्रान्तमा रहेको राष्ट्रिय निकुञ्ज	इन्डोनेसियाली पपुवा प्रान्तमा रहेको राष्ट्रि निकुञ्ज	0.167
3	चलचित्रमा केमियो रोलमा नायक राजबल्लभ कोइरालालाई पनि देख्न पाइनेछ	चलचित्रमा केमियो रोलमा नायक राजबल्लभ कोइराँलालाई पनि देख्न पाइनेछ	0.111
4	उनले दुई हजार दसमा जर्जियामा सुरु हुने स्टर्स टुर्नामेन्टमा भाग लिन थाले	उनले दुई हजार दसमा जर्जियामा सुरुहुने स्टोर्स टुर्नामेन्टमा भागलिन थाले	0.25

5 Conclusion

On performing several experiments it seem the performance of model depends upon several factors such as learning rate, number of epochs, momentum, batch size, training duration etc. The system predict the unseen data with the WER of around 11% which is quite satisfactory. It can be conclude that CNN-RNN architecture can be used for speech to teach conversion. The quality of the model depends upon the quality of the data. So before the training phase the data must kept clean by preprocessing on data. The model depends upon several factor. From the experiment it seem the batch size and learning rate greatly effect on model development. Several parameters must be hypertuned to obtain the best model. Finally it is concluded that CNN-GRU model can be implemented to develop Neplai ASR systm.

References

[1] Dong Liu, Antoine Honore, Saikat Chatterjee, and Lars K Rasmussen. Powering hidden markov model by neural network based generative models. *arXiv preprint arXiv:1910.05744*, 2019.

[2] Vishal Passricha and Rajesh Kumar Aggarwal. Convolutional neural networks for raw speech recognition. In *From Natural to Artificial Intelligence-Algorithms and Applications*. IntechOpen, 2018.

[3] Dong Wang, Xiaodong Wang, and Shaohe Lv. End-to-end mandarin speech recognition combining cnn and blstm. *Symmetry*, 11(5):644, 2019.

[4] Manish K Ssarma, Avaas Gajurel, Anup Pokhrel, and Basanta Joshi. Hmm based isolated word nepali speech recognition. In *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 71–76. IEEE, 2017.

[5] Paribesh Regmi, Arjun Dahal, and Basanta Joshi. Nepali speech recognition using rnn-ctc model. *International Journal of Computer Applications*, 178(31):1–6, Jul 2019.

[6] William Song and Jim Cai. End-to-end deep neural network for automatic speech recognition. *Standford CS224D Reports*, 2015.

[7] Hani S Matloub, David L Larson, Joan C Kuhn, N John Yousif, and James R Sanger. Lateral arm free flap in oral cavity reconstruction: a functional evaluation. *Head & neck*, 11(3):205–211, 1989.

[8] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016.

[9] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[10] Sadaoki Furui. 50 years of progress in speech and speaker recognition research. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, 1(2):64–74, 2005.

[11] Frederick Jelinek. The development of an experimental discrete dictation recognizer. *Proceedings of the IEEE*, 73(11):1616–1624, 1985.

[12] R Gupta and G Sivakumar. Speech recognition for hindi language. *IIT BOMBAY*, 2006.

[13] Adrien Meynard and Bruno Torrésani. Spectral analysis for nonstationary audio. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2371–2380, 2018.

[14] Sunil Kumar Kopparapu and M Laxminarayana. Choice of mel filter bank in computing mfcc of a resampled speech. In *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)*, pages 121–124. IEEE, 2010.

[15] Deividas Eringis and Gintautas Tamulevičius. Improving speech recognition rate through analysis parameters. *Electrical, Control and Communication Engineering*, 5(1):61–66, 2014.

[16] Chongchong Yu, Yunbing Chen, Yueqiao Li, Meng Kang, Shixuan Xu, and Xueer Liu. Cross-language end-to-end speech recognition research based on transfer learning for the low-resource tujia language. *Symmetry*, 11(2):179, 2019.

[17] Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. *arXiv preprint arXiv:1905.01997*, 2019.

[18] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[19] Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, and Yoshua Bengio. Light gated recurrent units for speech recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):92–102, 2018.

[20] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE, 2017.

[21] Xiang Kong, Jeung-Yoon Choi, and Stefanie Shattuck-Hufnagel. Evaluating automatic speech recognition systems in comparison with human perception results using distinctive feature measures. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5810–5814. IEEE, 2017.

[22] Keshan Sodimana, Knot Pipatsrisawat, Linne Ha, Martin Jansche, Oddur Kjartansson, Pasindu De Silva, and Supheakmungkol Sarin. A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 66–70, Gurugram, India, August 2018.