Proceedings of 10th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2021 Month: October Volume: 10

Nepali Speech Recognition using LSTM-CTC

Rupesh Shrestha [a], Basanta Joshi [b], Suman Sharma [c]

[a, b, c] *Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, Tribhuvan University, Nepal*
Corresponding Email: [a] 075MSICE018.rupesh@pcampus.edu.np, [b] basanta@ioe.edu.np

Abstract
Speech recognition system is developed which is able to transcribe an audio input to text format. The system is based on a combination of the Long short-term Memory neural network architecture and the connectionist temporal classification function. Word Error Rate of 40% for isolated word recognition without connectionist temporal classification layer and 34.3% on sentence recognition of Nepali speech corpus with connectionist temporal classification is achieved.

Keywords
Artificial Intelligence, Automatic Speech Recognition, Connectionist Temporal Classification (CTC), Softmax, Nepali Speech Recognition, Long Short-Term Memory (LSTM)

1. Introduction

Speech has been fundamental form of communication since the human civilization has begun. The Speech recognition could help in recognizing human speech and secondly communicate with the computers so as to ease the interactions of human with computer [1].

There have always been improvements in the how human interact with computers. It all began from the switching of the vacuum tube diodes and triodes that began how humans interacted with the computer. Most widely and accepted means of human–computer interface has been text. Texts are used for programming the computers and express human ideas, thoughts or simply to communicate with the computers. But as the computer science improved over the time computers began to be self-learning or simply saying machine learning got evolved. Additionally, it could help in transforming how humans interact with computer. This can be special help for disabled person who cannot operate computers and devices as normal human [2].

They can interact with computers in the form of speech. As the importance of speech has been realized in the task of expressing the human feelings and thoughts, speech recognition has been implemented in the modern machines by training them to recognize different speech components of the human speech. First Step is to extract useful features from the raw speech. Second step involves generating acoustic model where speech is represented as combination of different phones. Automatic Speech Recognition objective is to convert speech into its textual representation [3]. It involves converting textual features into components of text like- word, character or sentence. Various methods can be used to extract the temporal audio data. Probabilistic methods like-HMM/GMM had been used at the early stages of evolution of speech recognizer [4][5][6]. Invention of neural networks encouraged use of neural networks to memorize temporal speech features in form of recurrent neural networks. Some improvisation in recurrent neural networks to omit inherent defects like- gradient vanishing and bursting issue resulted in development of LSTM and GRU networks. Earlier development of speech recognition tools as mentioned in [2] has used MFCC feature for isolated word recognition and fail to recognize sentence with different speed of utterance and blank or spaces between words.

ASR basically consists of different blocks to perform sequence of tasks. First step is the pre-processing of the audio files like- noise removal. Later audio files are converted to various audio features or acoustic features. Development of speech recognition in Nepali language lag in time relative to English speech recognition. Research for Nepali speech recognition has helped in continuous development of Nepali ASR [7].

2. Methodology

Automatic Speech Recognition system requires series of steps and processes. Major steps in the ASR are Data preparation, feature extraction, model building, training, testing and validation. Major blocks considered for the Nepali speech recognition using LSTM-CTC algorithms has been shown in block diagram.

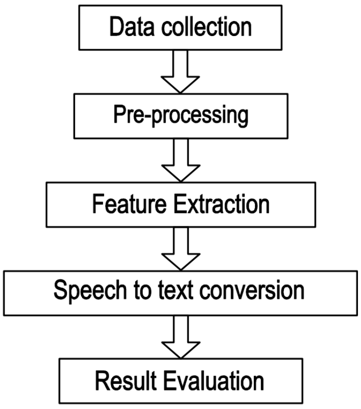


Figure 1: Block Diagram of Nepali Speech Recognition System using LSTM-CTC algorithm.

Clean dataset is preliminary requirement to use any machine learning algorithm. Dataset preparation is one of the major tasks during the research work involving machine learning algorithm.

Nepali Speech corpus that has total of 2813 audio speech files spoken by 3 different male speaker has been used to train, test and validate the Nepali speech recognition using LSTM-CTC. This Nepali speech corpus contains total 346 Nepali speech phrases spoken by different speakers for varying number of times. All audio files has been sampled at 16000 Hz with each sample represented by 16 bits. Among 3 speakers, audio speech of speech recording has low noise but we can hear some echo effects due to recording conditions in audio files by speaker 3 and worse for speaker 2.

After the dataset preparation data cleansing is carried out. The feature gets extracted and these features are passed into the neural networks for model to be trained [5].

Features are the individual measurable property or characteristic of a parameters being observed, that can be used as a distinctive attribute or aspect of something. The objective of the feature extraction is to perform classification more efficient and convenient.

Mel Frequency Cepstral Coefficients (MFCC) features are a sequence of MFCC Acoustic feature vectors where each vector represents information in a small-time signal window [3]. MFCC has been used for feature extraction from audio files.

Table 1: Summary of Nepali Speech Corpus.

Features	Speaker 1	Speaker2	Speaker 3
Sampling rate	16 KHz	16 KHz	16 KHz
BPS	16 bits	16 bits	16 bits
Bit rate	256Kbps	256Kbps	256Kbps
Utterance	Moderate	Slow	Slow
Echo	low	High	Moderate
Voice Clarity	Fine	Low	Moderate

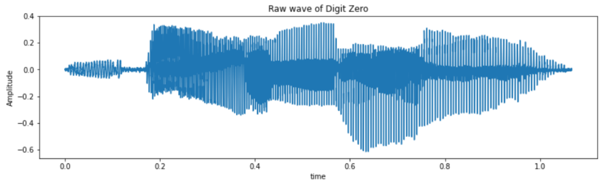


Figure 2: Time Domain Representation of sample audio signal.

MFCC shows the spectrum distribution for windowed audio signals. Sample MFCC plot has been shown.

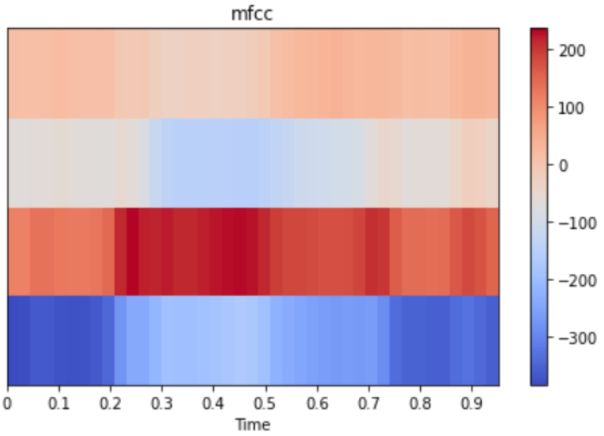


Figure 3: MFCC plot of sample audio signal.

Combination of Long Short-Term Memory (LSTM) and Bidirectional RNN has been used to generate Nepali transcripts corresponding to audio signals. LSTM helps in learning long-time sequence patterns. Bi-directional RNN output is processed by dense layer. And the output vector from dense layer is as of same size as total number of characters defined in

Nepali character sets with additional extra blank character. The output has then been forwarded to the SoftMax layer which assigns occurrence probability to each Nepali character. The probability vector thus obtained has forwarded to the CTC layer which again calculates CTC loss and has back propagated to reduce the loss.

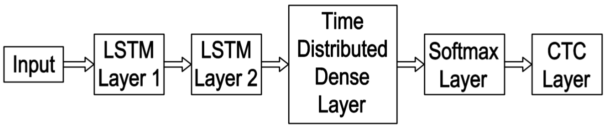


Figure 4: LSTM-CTC Model Architecture for Nepali Speech Recognition.

Incapability of learning long time-dependent patterns by standard RNN nodes compelled use of LSTM cells widely. Time sequence data has been trained as frame-wise classifiers in which the training dataset has target label for every frame. Similarly, Speech Recognition has similar situation where we need to match unequal frame size and character label. Same number of character labels might result in varying number of speech frames depending upon the reading speed of the reader, this makes alignment of speech frames and character labels very difficult. Connectionist Temporal Classifiers (CTC) helps in resolving this problem [1]. CTC loss is an error function that helps to train RNN or LSTM over time sequence data that does not have alignment among input speech frames and target character labels [5].

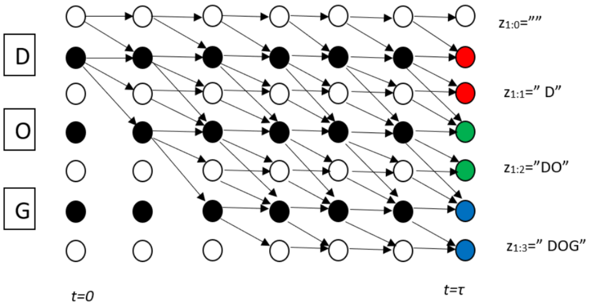


Figure 5: Working of CTC Algorithm to decode Nepali Speech in ASR.

The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level. The WER is a valuable tool for comparing different systems as well as for evaluating improvements within one system. This kind of measurement, however, provides no details on the nature of translation errors and further work is therefore required to identify the main source(s) of error and to focus any research effort.

This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment.

'क', 'ख', 'ग', 'घ', 'ङ', 'च', 'छ', 'ज', 'झ', 'ञ', 'ट', 'ठ', 'ड', 'ढ', 'ण',

'त', 'थ', 'द', 'ध', 'न', 'प', 'फ', 'ब', 'भ', 'म', 'य', 'र', 'ल', 'व', 'श', 'ष',

'स', 'ह', 'अ', 'आ', 'इ', 'ई', 'उ', 'ऊ', 'ए', 'ऐ', 'ओ', 'औ', ' ◌ा', ' ◌ि',

'◌ी ', '◌ु ', '◌ू ', ' ◌ृ', ' ◌े ', ' ◌ै ', ' ◌ो ', ' ◌ौ ', ' ◌्', ' ◌ँ ',

'०', '१', '२', '३', '४', '५', '६', '७', '८', '९','।', ' '

Figure 6: 67 Devanagari Characters used to write Nepali Texts.

The 67 Nepali character set is indexed from 0 to 66 representing '*ka*' as index: 0 and increasing accordingly. Now the output of dense layer is fed to SoftMax layer. The output of the Softmax work is identical to a categorical probability distribution, it gives the probability that any of the classes are true or not.

3. Results and Discussion

Nepali Speech Recognition tool has been developed using LSTM-CTC algorithm that can seamlessly convert the continuous Nepali sentence speech into frames with different audio characteristic features and later detect the Nepali texts that corresponds to the Nepali speech frames. The Nepali Speech Recognition was successfully reconstructed using this methodology.

LSTM has been used to convert speech features extracted using methods like MFCC to model the Nepali characters from the speech characters. Additionally, CTC algorithm has been used efficiently for Nepali speech recognition for varied speed of Nepali word utterance. Furthermore, the set of 67 Nepali characters has been defined as sufficient set of Nepali characters to transcribe Nepali speech database taken as dataset even if there are total of 128 possible Nepali characters possible. In order to evaluate working efficiency of CTC algorithm used in decoding, Performance of the ASR model without CTC layer has been used to recognize isolated words and sentence.

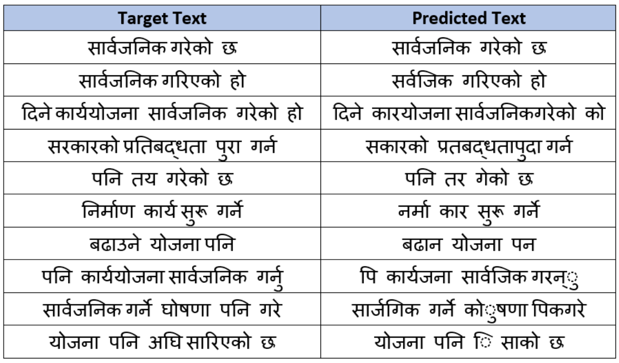
Table 2: WER comparison of isolated word and sentence recognition.

Speech Type	WER(%)
Isolated Word Recognition	40%
Sentence Recognition	> 70%

Even without CTC ASR results for isolated word recognition was satisfactory and gave WER of about 40%. However, evaluation of Automatic Speech Recognition without CTC layer shows that it is not preferred to be used for the sentence recognition as we need to incorporate blank or spaces in sentence. Performance of the ASR without CTC layer is not satisfactory even for single word detection due to varying speed of utterance of words by different speakers.

Later, CTC layer was added to the ASR model making it capable to detect words spoken at different speed. This made ASR model capable for sentence level recognition. ASR model has now been trained by Nepali Speech corpus. Training and validation loss observed over at different epochs has been gradually stabilized.

Some results of Nepali speech recognition using LSTM-CTC algorithm has been demonstrated.



Target Text	Predicted Text
सार्वजनिक गरेको छ	सार्वजनिक गरेको छ
सार्वजनिक गरिएको हो	सर्वजिक गरिएको हो
दिने कार्ययोजना सार्वजनिक गरेको हो	दिने कारयोजना सार्वजनिकगरेको को
सरकारको प्रतिबद्धता पुरा गर्न	सकारको प्रतबद्धतापुदा गर्न
पनि तय गरेको छ	पनि तर गेको छ
निर्माण कार्य सुरू गर्ने	नर्मा कार सुरू गर्ने
बढाउने योजना पनि	बढान योजना पन
पनि कार्ययोजना सार्वजनिक गर्नु	पि कार्यजना सार्वजिक गरन्ु
सार्वजनिक गर्ने घोषणा पनि गरे	सार्जगिक गर्ने कोुषणा पिकगरे
योजना पनि अघि सारिएको छ	योजना पनि ि साको छ

Figure 7: Some very accurate predictions by Nepali ASR.

This shows that the model works very well for some of the speech segment but could not perform proper prediction for the rest of the Nepali speech datasets. As for testing, while isolated word dataset with 10 epochs to fit the model on generated network. The test generated WER of 60%.

Validation of the thus built Speech recognition model for isolated word was performed on the audio files recorded in normal room condition. The validation of the model results was only 30% spoken words would be recognized by ASR model.

Nepali speech recognition tool is DNN with LSTM and CTC implemented using different tools like keras and librosa available for python coding platform. Available audio datasets has been split into the training, test and validation datasets. The spoken isolated words consisting of recordings of spoken isolated words sampled at 8kHz which are available in '.wav' format and have fixed bit rated of 64kbps after representing each sample by 8bits. The audio clips has been obtained and labelled as per the dataset descriptions.

Secondly, for the testing of the codes for Nepali speech recognition, we provide ASR model with this training data and interconnect the neurons with optimized learning rate and observe the accuracy for different Epochs. Average Word Error Rate (WER) of the 58.97% was observed for model trained over 20 epochs. And average WER of 9.62% was observed for model trained over 30 epochs.

Table 3: WER observed for different speakers for different values of Training Epochs.

Speakers	WER(%)	
	Epoch=20	**Epoch=30**
Speaker 1	40.00%	2.27%
Speaker 2	66.93%	17.60%
Speaker 3	70.00%	9.00%
Average	**58.97%**	**9.62%**

ASR model for Nepali Speech Recognition has been developed using Dense Neural Network (DNN). The developed model has been used to predict Nepali Speech data and generated the corresponding Nepali texts. As Nepali text involves many characters, noisy training datasets and due to low efficiency of the model very high WER was observed. As shown in the table, accuracy for speech recognition is improved for the audio that have clear voice and lesser disturbances due to echo. For the model built with 20 epochs of training WER is comparatively high. Speaker 1 have clearest of voice gave lowest error of 40% while the audio from speaker 2 and 3 who had noisy and unclear voices result in high error rates of 66.93% and 70% respectively.

The ASR model using LSTM-CTC for Nepali speech recognition trained for 30 epochs resulted in better accuracy. The effect of noise and voice clarity has been factor affecting accuracy in this case too. Due to higher training epoch, overall accuracy has been reduced to 9.62%.

4. Conclusion

DNN developed using LSTM and CTC algorithm recognized the Nepali text corresponding to the Nepali speech input data. Average Word Error Rate (WER) of the 58.97% was observed for model trained over 20 epochs. And average WER of 9.62% was observed for model trained over 30 epochs. WER has been degraded in case of voice deteriorated with echo or noise and fast speakers. Thus, clear voice, slower speech and low noise recording environment data results ASR that gives lower WER.

Acknowledgments

This work has been supported by the University Grants Commission, Nepal under a Faculty Research Grant (UGC Award No. FRG-76/77- Engg-1) for the research project "Preparation of Nepali Speech Corpus: Step towards Efficient Nepali Speech Processing".

References

[1] Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR, 2014.

[2] Chetan Prajapati, Jiwan Nyoupane, Jwalanta Deep Shrestha, and Shishir Jha. Nepali speech recognition. *Kathmandu. DOECE*, 2008.

[3] Ripul Gupta. Speech recognition for hindi. *M. Tech. Project Report, Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Mumbai, India*, 2006.

[4] 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.

[5] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2011.

[6] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[7] A Kalakheti, KP Bhattarari, S Kuwar, and S Adhikari. Automatic speech recognition for nepali language. *Tribhuvan University, Nepal*, 2013.