Bidirectional Long short-term memory and Recurrent Neural Network model for speech recognition

Mercy Kimani[1], Lawrence Nderu[2], Dalton Ndirangu[3], and Mwalili Tobias[4]

[1]Machakos University
[2]Jomo Kenyatta University of Agriculture and Technology
[3]United States International University School of Science and Technology
[4]Jomo Kenyatta University of Agriculture and Technology College of Pure and Applied Sciences

July 5, 2023

Abstract

Speech-to-text is essential as it converts spoken words to text, thus making it easy to store. It has several components; from a basic model, it is viewed in four stages; Signal pre-processing, feature extraction, feature selection, and modeling. Several works of literature have been documented on improving and achieving better results in speech recognition. However, works remains in resolving the issue of word error rate and accuracy on continuous input stream without increasing the required bandwidth. This research evaluates recurrent neural networks, long short-term memory neural networks, gated recurrent units, and bi-directional long short-term memory. It further tests the signal's performance after introducing bias to the long short-term memory. This research then proposes a model bi-directional long short-term memory recurrent neural network. Experimental results demonstrate that even with a bias of one on long short-term memory, the bidirectional long short-term memory recurrent neural network model still achieves better results with a word error rate of 8.92%, accuracy of 91.08% and mean edit distance of 0.1910 using the Libri speech training dataset. Future work will evaluate the use of the transformer models in the reduction of the word error rate and accuracy on a continuous input stream.

Bidirectional Long short-term memory and Recurrent Neural Network model for speech recognition.

Mercy Wairimu Kimani[1] | Dr Lawrence Nderu[2] | Dr Dalton Ndirangu[3] | Dr Tobias Mwalili[41]School of Engineering and Technology, Machakos University, Machakos, Kenya[2] School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya[3]School of Science and Technology, United States International University Nairobi, Kenya[4]School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya**Correspondence**Mercy Wairimu Kimani, School of Engineering and Technology, Machakos University, PO Box 136-90100 Machakos, Kenya.

Email: Mercyk630@gmail.com

mercy.wairimu@mksu.ac.ke

Abstract

Speech-to-text is essential as it converts spoken words to text, thus making it easy to store. It has several components; from a basic model, it is viewed in four stages; Signal pre-processing, feature extraction, feature selection, and modeling. Several works of literature have been documented on improving and achieving better results in speech recognition. However, works remains in resolving the issue of word error rate and accuracy

on continuous input stream without increasing the required bandwidth. This research evaluates recurrent neural networks, long short-term memory neural networks, gated recurrent units, and bi-directional long short-term memory. It further tests the signal's performance after introducing bias to the long short-term memory. This research then proposes a model bi-directional long short-term memory recurrent neural network. Experimental results demonstrate that even with a bias of one on long short-term memory, the bidirectional long short-term memory recurrent neural network model still achieves better results with a word error rate of 8.92%, accuracy of 91.08% and mean edit distance of 0.1910 using the Libri speech training dataset. Future work will evaluate the use of the transformer models in the reduction of the word error rate and accuracy on a continuous input stream.

KEYWORDS

Speech recognition, Word Error Rate, Accuracy, Bi-directional LSTM, Recurrent Neural Networks

Introduction

Speech recognition is the conversion of the spoken word into text, enabling the creation and use of speech information. Speech-to-text is a crucial application as it is easy to store text, however indexing a specific utterance can be difficult as speech signals can be swift, intuitive, slow, and unpredictable at other times (Gupta & Joshi, 2018). Over the years, speech processing has evolved with the current state-of-the-art systems replaced. According to (Nassif et al., 2019), traditional speech recognition systems work by encoding the speech impulses using the Gaussian Mixture Model (GMM) which is based on Hidden Markov Model (HMM) as it can be viewed as a stationary signal. Dynamic Time Warping (DTW) algorithm, has been used for speech recognition and though it has been replaced with other new algorithms it's still a very popular technique. This algorithm works very well by identifying the optimum distance which is the shortest distance between two speakers therefore addressing the issue of speed between speakers. In general speech recognition systems can be viewed in four stages: signal preprocessing, feature extraction, feature selection, and modeling.

Automatic speech recognition using Deep Neural Networks has grown in popularity. This technology typically involves categorizing acoustic templates into pre-established classes. Numerous studies have provided numerous instances illustrating how deep neural networks do better than conventional models at speech recognition. The final results of Microsoft Audio Video Indexing Service, a deep learning-based speech system that was disclosed by Microsoft, showed a 30% reduction in word error rate (WER) on four benchmarks when compared to the state of models based on Gaussian mixtures. (Samal, Jena, & Manjhi, 2019). Time-series data are involved in the speech recognition issue.

Deep neural networks are constructed in a series of layers, each of which contains neurons that take input from the layer above and conduct a single computation. The outputs of the previous levels are sent to the subsequent layer in feedforward neural networks because they are unidirectional. The feedforward networks' inability to transmit historical information is one of their drawbacks. Additionally, problems like varying speaking rates and temporal dependencies frequently arise when Deep Neural Networks are employed to analyze speech recognition. By modeling a fixed sliding window of auditory frames, Deep Neural Networks are comfortable and able to handle temporal dependencies, but they are unable to accommodate varying speaking rates. Recurrent neural networks (RNN) are a subset of artificial neural networks where the connections between the nodes have the potential to cycle because they have loops in the hidden layers that store information from previous layers, allowing the output from one node to influence the subsequent input to that same node and predicting the value of the current step. Due to this RNN can handle the difficulty of diverse speaking rates because of this method (Apeksha Shewalkar, 2019).

Connectionist Temporal Classification (CTC) is a type of neural network that can resolve the limitations of RNN that require pre-segmented training as well as post-processing of the output given to convert it into labeled sequences. Long Short-Term Memory (LSTM) networks address the issue of long-term dependencies in the data which made the networks very effective. Gated Recurrent Unit (GRU) Neural Networks work well with sequential data thus also effectively handling the issue of long-term dependencies. Although integrated

models have been developed to address the issues of accuracy and speed in speech recognition, insufficient details have been focused on how a model can be integrated to address all the issues raised by the various algorithms proposed. In this study, we integrate, BLSTM and recurrent neural networks to address the highlighted challenges specifically the continuous speech input stream.

In recent years, transformer architectures based on self-attention mechanisms have proved to demonstrate the ability to outperform Bidirectional Long-Short Term Memory (BLSTM) giving good results on acoustic modeling (Yongqiang Wang, 2020). With the introduction of transformers, gaps identified such as the limits of length of speech signals that are required for long-term range dependencies were addressed, and the fact that recurrent models do not allow parallelization transformers can address this challenge.

Although integrated models have been developed to address the issues of accuracy and speed in speech recognition, insufficient details have been focused on how a model can be integrated to address all the issues raised by the various algorithms proposed.

In this study, we integrate, BLSTM and recurrent neural networks to address the highlighted challenges specifically the continuous speech input stream.

Contributions of this study: -

1. Through investigations of several architectures in LSTM, GRU, and RNN this paper proposes an enhanced model Bidirectional LSTM recurrent Neural Network (BLSTM-RNN).

2. To demonstrate the superiority of the proposed model in resolving the issue of word error rate and accuracy on continuous input stream without increasing the required bandwidth

Proposes an integrated model based on DTW, BLSTM, and transformers for speech recognition.

The paper is structured as follows. Section 2 discusses related work using various in the area of speech recognition. Section 3 discusses the proposed approach. Section 4 describes the datasets and experimental environment. Section 5 discusses the results. The conclusion and future work are provided in section 6.

Literature Review

From reviewed literature, work has been documented on ways to improve and achieve better results in speech recognition. (Hori et al., 2017) developed a model that was based on End-to-End encryption using advances in Joint CTC-Attention for Speech Recognition with a Deep CNN Encoder and RNN-LM. In their work, they combined the CTC predictions, the attention-based decoder prediction, and a separately trained LSTM language model. This research achieved a 5-10% error reduction compared to prior systems on spontaneous Japanese and Chinese speech and other traditional Automatic Speech recognition systems. For future studies, this work recommended improvement on the model using vast quantities of unlabeled data to pre-train RNN-LM which could be jointly trained with the recommended model.

(Kumar et al., 2018) Provided a survey on deep learning techniques and the architectures involved in speech recognition. In this paper, different models were discussed based on Deep Belief Networks, Recurrent Neural Networks, and Convolutional Neural networks. From their findings, DBNs and CNNS have been demonstrated to work very well with large vocabulary replacing the Gaussian mixtures. In addition, to achieve end-to-end speech recognition, the proposed elimination of processing stages and instead recommended the use of one unified neural network. In their conclusion findings, the usage of Recurrent Neural Networks for acoustic modeling in a hybrid Dynamic Neural Network Hidden Markov Model had mixed reactions using the CTC loss function and language model. Deep learning, therefore, holds the power to deal with raw inputs eliminating difficult processing stages and learning-rich representations.

(James1 et al., 2018) did a study on end-to-end speech recognition using LSTM networks for electronic devices, the model demonstrated promising performance with a Word Error Rate reduction of 6.8% as compared to the hybrid model and WER of 11.9% from the baseline system. The enhanced performance of the model is contributed by the exploitation of modeling capacity for LSTM to directly process speech

signals and exploit past information to reliably estimate speech parameters. The results obtained in this study demonstrate the promising potential of the LSTM model as a renowned technique in continuous speech recognition and utilized to be implemented as a hardware model for controlling stand-alone electronic devices.

(Mokgonyane et al., 2019) developed an automatic speech recognition system that incorporated four classifier models used in machine learning support vector machine (SVM), K-nearest neighbors (KNN), Random Forest (RF), and Multi-layer Perceptron (MLP). To determine the best classifier, the researchers applied the auto Weka data mining tool with its best hyperparameters. To evaluate the performance of the model, 10-fold cross-validation was used. In conclusion, models trained compared to CNN, DNN, and LSTM models incorporating several parameters, and the experimental results showed that the LSTM model performs better in the test data and training data. Noise reduction was best achieved from 31.23% to 25.89%. This majorly informs the researcher's interest in conducting further tests on recurrent neural networks and bidirectional long short-term memory.

(Daneshfar & Kabudian, 2020) developed a speech recognition model using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. The results of applying the method on large emotional datasets such as the Berlin Database of emotional speech (EMO-DB), Surrey Audio-Visual Expressed Emotion (SAVEE), and Interactive Emotional Dyadic Motion Capture (IEMOCAP) showed that in terms of accuracy, the proposed pQPSO algorithm outperformed standard QPSO algorithm, wQPSO. Classical dimensionality reduction methods, deep neural networks, and state-of-the-art methods on the same datasets. In addition to optimizing the GMM parameters and the transformation matrix for dimensionality reduction, other applications of the pQPSO algorithm include optimizing the MFCC filter bank parameters, optimizing the classifier parameters, and finding more features related to emotion.

(Koduru et al., 2020) did a study on Feature extraction algorithms to improve the speech emotion recognition rate. The research found out that from the simulation results, they depicted the accuracy and efficiency of different classifiers that achieved an accuracy of Support Vector Machine of 70%, Decision tree is 85%, and LDA of 65%. By comparing the results, the proposed system achieved more accuracy than the existing work implying that it extracts maximum information from the signal required to represent the characteristics and recognize emotions from the signal efficiency.

(Ho et al., 2020) in their study on Prediction of Time Series Data Based on Trans-former with Soft Dynamic Time Wrapping using open-source dataset House Twenty shows that the average prediction error rate with soft-DTW Transformer is 27.79%, greatly reduced from 45.70% for using SVR, a common time series method. The research recommended more experiments would be conducted to compare with other methods.

(Gulati et al., 2020) conducted a study on a convolution-augmented transformer for speech recognition known as conformer. This research demonstrated that the developed model conformed significantly outperformed the previous transformer and CNN-based models this achieving state of the are accuracies. The study used the widely used Libri speech benchmark dataset and was able to achieve WER of 2.1%/4.3% without using a language model and 1.9%/3.9% with an external language model on test/test other. We also observe the competitive performance of 2.7%/6.3% with a small model of only 10M parameters.

(Pawar & Kokate, 2021) in their work developed a Convolutional neural network (CNN) based on emotion recognition using Mel-frequency Cepstrum coefficients. In the study, CNN architecture working with deep learning behavior provided the best classification of accuracy by reducing the complexity of speech with the help of combined feature extraction algorithms such as Pitch and Energy, Mel-Frequency Cepstral Coefficients (MFCC), and Mel Energy Spectrum Dynamic Coefficients (MEDC) and selection methods. In comparison with the existing methods like KNN classifiers, CNN was found to perform better and provided the best results in terms of ROC and AUC characteristics curves. In the research, they recommended the use of proper classification and feature extraction methods for better accuracy in emotion recognition.

Methodology

The dataset

The dataset used for the research is the LibriSpeech ASR corpus which includes 1000 hours of recorded speech. The latest dataset has an improved language model which is an important factor in achieving reduced WER (Word Error Rate) values. It is publicly available and is designed to train acoustic models and language models. It is separated into data folders for training, validation, and testing. It consists of 16kHz audio files between 2-15 seconds long of spoken English derived from read audiobooks from the LibriVox project. For this experiment, the audio files were converted to single channel (mono) WAV/WAVE files (.wav extension) with a 64k bit rate, and a 16kHz sample rate and then encoded in PCM format, and then cut/padded to an equal length of 10 seconds. The pre-processing techniques used for the text transcriptions included the removal of any punctuation other than apostrophes, and transforming all characters to lowercase. In total, there are 64220 total training examples.

Experimental Set-up

Experimentation was done under the Google Cloud compute engines custom machine which is accessed with a personal laptop computer. Python 3.6 from Continuum Analytics Anaconda has been used to set up to create a virtual environment and sci-kit-learn, Tensor flow, Keras, and graph lab create libraries configured to work in jupyter notebook. Jupyter Notebook was accessed through the browser and all developed Python code has been typed and executed. Matlab software assisted with the algorithm development.

Evaluation Measures.

In speech recognition, there are two different types of performance or evaluation measures, which are based on (1) accuracy, and (2) speed. Evaluation measures based on accuracy include WER and mean edit distance.

Word Error rate (WER) is calculated as: -

$$WER = \frac{S + I + D}{N} * 100$$

Where: -

S is the number of substitutions,

I is a number of insertions,

D is the number of deletions,

N is the total number of words in the actual transcript.

The interpretation of WER is that the lower the WER, the better the speech recognition

Mean Edit Distance - is a measurement of how many changes we must do to one string to transform it into the string we are comparing it to. Let us say the normalized edit distance between two words/strings (consider A and B) is d (A, B). The mean edit distance is calculated by: -

$$d\,(A, B) = min(\frac{W\,(P)}{N})$$

Where: -

where P is the editing path between string A and string B

W (P) is the total sum of weights of all the edited operations of P

and N is the total number of edited operations (the length of the editing path, P)

Parameter setup

The following parameters were used to run the model: -

- Dropout rate = 30%
- Number of epochs = 20
- Training batch size = 16
- Test batch size = 8
- Activation function = ReLU
- Neuron count in hidden layers = 1,000
- Adam optimizer:
- β1 = 0.9
- β2 = 0.999
- ε = 1e-8
- learning rate = 0.0001
- **Experiment Process flow**
- **Model Architecture**

Results and Analysis

The experiments were conducted on five models: RNN, GRU, LSTM, Bidirectional LSTM, LSTM with bias, and Bidirectional LSTM-RNN model. The models were run on 1000 node architectures in each hidden layer. Table 1.0 provides the results of the WER, RNN achieved 94.34% on the training dataset, 94.05% on the validation data set, and 94.31 on the testing dataset. GRU achieved 29.95% on the training dataset, 94.05% 0n the validation dataset, and 94.31% on the testing dataset. LSTM achieved 28.30% on the training dataset, 31.76% on the validation dataset, and 32.82 on the testing dataset. BLSTM achieved 19.92% on the training dataset, 25.07% on the validation dataset, and 26.26% on the testing dataset. LSTM with bias initialized to one achieved 19.92 on the training dataset, 33.67 on the validation dataset, and 34.42 on the testing dataset. Bidirectional LSTM-RNN achieved 8.92% on the training dataset, 11.46% on the validation dataset, and 13.07% on the testing dataset. BLSTM-RNN model scoring best. Table 1.1 provides results of the model accuracy simple RNN achieved 5.66% and training dataset 5.95% on the validation dataset and 5.69 on the testing dataset. GRU achieved 70.05% on the training dataset, 66.37% on the validation dataset, and 65.58% on the testing dataset. LSTM achieved 71.7% on the training dataset, 68.24% on the validation dataset, and 67.18% on the testing dataset. BLSTM achieved 80.08% on the training dataset, 74.93% on the validation set, and 73.74% on the testing dataset. LSTM with bias initialized to one achieved 61.6% on the training dataset, 58.18% on the validation dataset, and 57.16% on the testing dataset and Bidirectional LSTM-RNN achieved 91.08% on the training dataset, 88.54% on the validation and 86.93% on the testing dataset, with BLSTM-RNN scoring best. In terms of the mean edit distance as shown in table 1.2 BLSTM-RNN model achieved the best value. Table 1.0 Word Error Rate.

Model	**Training dataset (WER %)**	**Validation dataset (WER %)**	**Testing dataset (WER %)**
LSTM	28.30	31.76	32.82
LSTM bias 1	38.40	41.82	42.84
BLSTM	19.92	25.07	26.26
GRU	29.95	33.63	34.42
Simple RNN	94.34	94.05	94.31
BLSTM-RNN	8.92	11.46	13.07

Table 1.1 Accuracy of the model

Model	**Training dataset (Accuracy %)**	**Validation dataset (Accuracy %)**	**Testing dataset (Accurac**
LSTM	71.7	68.24	67.18
LSTM bias 1	61.6	58.18	57.16

Model	Training dataset (Accuracy %)	Validation dataset (Accuracy %)	Testing dataset (Accurac
BLSTM	80.08	74.93	73.74
GRU	70.05	66.37	65.58
Simple RNN	5.66	5.95	5.69
BLSTM-RNN	91.08	88.54	86.93

Table 1.2 Mean Edit Distance (MED).

Model	Training dataset (MED)	Validation dataset (MED)	Testing dataset (MED)
LSTM	0.3739	0.3308	0.3221
LSTM bias 1	0.4333	0.3899	0.3566
BLSTM	0.3942	0.3879	0.3556
GRU	0.5662	0.6443	0.4566
Simple RNN	0.7881	0.6882	0.4556
BLSTM-RNN	0.1910	0.2312	0.2000

Discussion

The results indicate that the ensembled model Bidirectional LSTM-RNN model scored the best results on the Word Error rate, accuracy, and Mean Edit Distance (MED). Parameter optimization of the experiment setup influenced the better performance as highlighted by (Apeksha Shewalkar, 2019). This research adopted the random search hyperparameter optimization method. While previous research has focused on performance based on different architectures, these results demonstrate that with parameter optimization good values can be achieved within acceptable running time days. In this research the number of hidden layer nodes architecture was limited to 1000 nodes this can be in the future increased to see the impact on the performance of each model.

Conclusion

This section has evaluated Gated Recurrent Unit (GRU), Simple Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), LSTM with a bias, BLSTM, and BLSTM-RNN model and compared the performance using LibriSpeech data set. The performance evaluation measure used was Word Error Rate, Mean Distance Edit, and accuracy as well as the training loss versus the validation loss as per the dataset. The results show that the BLSTM-RNN model performed best as compared to the other models. In future work more tests can be conducted on the same models on different architectures and the parameters too can be modified to see how the models will perform.

References

Guha, R., Ghosh, M., Kapri, S., Shaw, S., Mutsuddi, S., Bhateja, V., & Sarkar, R. (2021). Deluge based Genetic Algorithm for feature selection. Evolutionary Intelligence, 14(2), 357–367. https://doi.org/10.1007/s12065-019-00218-5 Gupta, A., & Joshi, A. (2018). Speech Recognition Using Artificial Neural Network. Proceedings of the 2018 IEEE International Conference on Communication and Signal Processing, ICCSP 2018, 333031, 68–71. https://doi.org/10.1109/ICCSP.2018.8524333 Hori, T., Watanabe, S., Zhang, Y., & Chan, W. (2017). Advances in Joint CTC-Attention based End-to-End Speech Recognition with a Deep CNN Encoder and RNN-LM. http://arxiv.org/abs/1706.02737 Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. International Journal of Speech Technology, 23(1), 45–55. https://doi.org/10.1007/s10772-020-09672-4 Mokgonyane, T. B., Sefara, T. J., Modipa, T. I., Mogale, M. M., Manamela, M. J., & Manamela, P. J. (2019). Automatic Speaker Recognition System based on Machine Learning Algorithms. Proceedings - 2019 Southern African Universities Power Engineering Conference/Robotics and

Mechatronics/Pattern Recognition Association of South Africa, SAUPEC/RobMech/PRASA 2019, 141–146. https://doi.org/10.1109/RoboMech.2019.8704837 Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech Recognition Using Deep Neural Networks: A Systematic Review. IEEE Access, 7, 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880 Pangotra, A. and others. (2020). Review On Speech Signal Processing & Its Techniques. European Journal of Molecular \& Clinical Medicine, 7(7), 3049–3052. Pawar, M. D., & Kokate, R. D. (2021). Convolution neural network-based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-020-10329-2 Shah, V. H., & Chandra, M. (2021). Speech Recognition Using Spectrogram-Based Visual Features. In S. E.-Y. X.-S. ED - Patnaik (Ed.), Advances in Machine Learning and Computational Intelligence (2021st ed., Vol. 1, pp. 695–704). Springer Singapore. Kunasekaran, K. K. H. (2015). Proceedings of the International Conference on Interdisciplinary Research in Electronics and Instrumentation Engineering 2015: ICIREIE 2015. Association of Scientists, Developers, and Faculties (ASDF)

Acknowledgments

1. Scholarship funders - International Development Research Centre (IDRC) and Swedish International Development Cooperation Agency (SIDA)
2. Scholarship Programme- Artificial Intelligence for Development (AI4D) Africa
3. Scholarship Fund Manager- Africa Center for Technology Studies (ACTS)

Conflict of Interest

Authors have no conflict of interest relevant to this article.