INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Use of Bidirectional Long Short Term Memory in Spoken Word Detection with reference to the Assamese language

Deepjyoti Kalita[1]*, Khurshid Alam Borbora[2], Dipen Nath[3]

1 Dept. of Computer Science & IT, Mangaldai College, Assam, India
2 Dept of Computer Science, IDOL, Gauhati University, Assam, India
3 Dept. of Computer Science, Kokrajhar Govt. College, Kokrajhar, Assam, India

OPEN ACCESS

Received: 22-03-2022
Accepted: 06-06-2022
Published: 18-07-2022

Citation: Kalita D, Borbora KA, Nath D (2022) Use of Bidirectional Long Short Term Memory in Spoken Word Detection with reference to the Assamese language. Indian Journal of Science and Technology 15(27): 1364-1371. https://doi.org/10.17485/IJST/v15i27.655

* **Corresponding author**.

deepjyoti111@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives : The proposed method is based on a unique technique of Deep learning for identifying spoken words with reference to Assamese language. Most of the DNN based algorithms have been successfully implemented in the field of image recognition, computer vision, natural language processing and medical picture analysis. **Methods:** The method used here is the Bidirectional Long Short Term Memory (BLSTM). BLSTM incorporates both past and future situations together. The speech database for this research work is hired from the repository of Indian Language Technology Proliferation and Development Center (ILTP-DC). This repository contains 32,335 utterances by 1000 numbers of male and female participants, which is comprised of 262 unique Assamese native words. The BLSTM based recognition model is using 10 out of the 262 unique words and the remaining words are used in construction or generation of synthesized sentences. The feature extraction module uses 39 feature coefficients, which are composed of MFCC, ΔMFCC and ΔΔMFCC coefficients. **Findings:** The Word Error Rate (WER) of the BLSTM based recognition model is 18.84% with an average accuracy of 98.12%, which sets one promising benchmark when compared to recent findings. **Novelty:** In this work an attempt has been made with a different approach to detect certain keywords of Assamese language by adopting deep learning methodology. The future objective of this proposed work is to improve the detection capability of this model by considering multiple DNN models together in a hybrid approach along with the inclusion of additional features.

Keywords: Bidirectional Long Short Term Memory; Deep Learning; Speech recognition; WER; MFCC

1 Introduction

Speech recognition aims to create intelligent systems capable of identifying and understanding the meaning of almost all words spoken by any speaker in any environment. It is the robustness of the system that determines its public acceptance. The speech recognition systems can assist us in basic tasks such as providing the driver

with a driving route, dialing a phone number, managing the computer by voice command, etc. Speech recognition may also assist physically challenged people by converting text to speech and vice versa [1,2]. A significant increase in the area of ASR-related research is noticed in the last two decades. Even though most of the other languages are used in voice recognition technology with steady improvements, the Assamese speech recognition has received very little attention. This is happening because of the structural complexity of the language and the scarcity of complete data on which to train any algorithm. In this research, a novel attempt has been made to conduct an experiment that detects specific spoken keywords from a set of news files. Many academicians have so far utilized a variety of keyword detection techniques in their works. A deep learning-based model (BLSTM) has been presented here, a form of a recurrent neural network, containing the most common aspects of a speech signal, MFCC, ΔMFCC and ΔΔMFCC for spoken word detection uttered by individuals in various conversation-based audio files. The length of the utterances fluctuates in ASR, which sets it apart from other sequence learning challenges. The same word may be uttered several times for varying amounts of time [3].

The most frequent approach to training a model for recognition is the Hidden Markov Model (HMM). HMMs bear the ability to accommodate variable durations utilizing Dynamic Temporal Warping (DTW). In languages, all words are made up of a small number of phonemes. HMM takes use of this as well, modeling the acoustic properties and a small number of phonemes accurately. The problem with classic speech recognition is the uniform and independent distribution of the observations.

Deep neural networks are revealed to offer a better approximation of posterior probability than the Gaussian mixture modeling (GMM). A hybrid model that integrates additional models like Deep model, HMM model, and singular value decomposition is also found effective [4]. Audio sequences may also be learned using recurrent neural networks. They frequently connect to memory storage devices. As a result, HMM states its clear advantages over the traditional ways of processing audios. The gates of the LSTM, which is a typical recurrent network, are used to store and retrieve memory. LSTM can be used to detect the phonemes present in individual words more accurately and efficiently. LSTM may also be convolved for end-to-end ASR models, and it produces superior outcomes in terms of Word Error Rate (WER) [5],[6],[7],[8].A similar model was proposed by D Kalita et al., 2019 considering AANN and MFCC and reported an accuracy percentage up to 87% [1].

Various RNN and CNN based methods have been applied to keyword spotting in recent years and impressive results are reported. Tang et al. [9] have employed Res15 to develop a neural network based keyword spotting system model with 95.8% accuracy on the Google Speech Commands Dataset (v1). Choi et al. [10] have used temporal convolutions and ResNet to build TC-ResNet models with 305K parameters, enhancing the accuracy percentage to 96.6%. Mittermaier et al. [11] have employed SincNet's parameterized Sinc-convolutions to categorize the keywords based on raw audio, reducing the number of parameters to 122K while preserving the TC-accuracy [12].

As per the literature study, deep learning methods are likely to perform well in Assamese spoken words because these methods can model speech signals efficiently in other Indian languages.

1.1 Long Short Term Memory (LSTM)

LSTM is a type of RNN-based deep learning architecture that is thought to be one of the most effective gradient-based methods. Unlike standard feed-forward neural networks, LSTM facilitates feedback connections. LSTM can handle single data points and whole data sequences (such as images, speech or video). LSTM can be used to connect handwriting recognition, speech recognition, detect anomalies in network traffic, and make intrusion detection systems. LSTM units are usually made up of a cell, an input gate, an output gate, and a forget gate (Figure 1). The values are saved in the cells for an unlimited amount of time. These are capable of learning order dependency, which is extremely valuable in sequence prediction problems. This is essential in various complex situations such as machine translation, speech recognition, and so on. RNNs are not the same as traditional feed-forward neural networks. Internal states of recurrent networks can be used to encode context information. RNNs remember prior inputs for a set length of time given by the weights and input data. An input sequence can be transformed into an output sequence using a recurrent network comprising non-fixed inputs that accommodates contextual information. RNN relies on context for its predictions, which it has to learn to some extent. Cycles in RNNs are used to impact current time step predictions based on information from prior time steps. In the internal states of the network, these activations are retained and may give long-term temporal context data. As the input sequence history progresses, RNNs can take use of a dynamically changing contextual window [13],[14].

LSTM contains three gates and they are mathematically represented as follows-

$$i_t = \sigma(x_t U^i + h_{t-1} W^i)$$

$$f_t = \sigma\left(x_t U^f + h_{t-1} W^f\right)$$

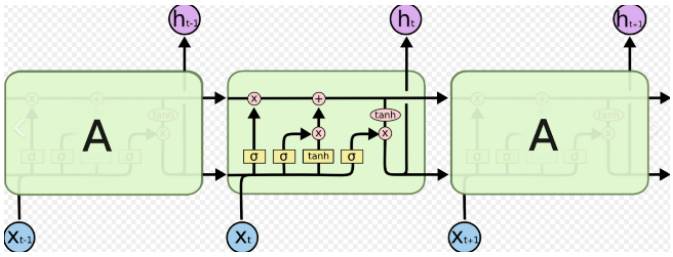

Fig 1. The structure of LSTM Model

$$o_t = \sigma(x_t U^o + h_{t-1} W^o)$$

1.2 Bidirectional LSTM

A bidirectional RNN (BRNN) is a model created to lessen some of the limitations of standard RNNs. This design divides recurrent networks into two categories, viz., forward and backward recurrent networks. These two networks connect to the same output layer to generate the output data. With this framework in place, successive inputs' past and future circumstances may be examined without delay. The LSTM equivalent of the BRNN structure is bidirectional LSTM (BLSTM), whose design model is depicted in Figure 2 . The LSTM model's performance in classification operations may be enhanced with this version. Unlike the typical LSTM structure, each job is trained with two separate LSTM networks in BLSTM. The LSTM networks at the bottom demonstrate the forward features. The networks mentioned above are used backward. Both networks use a single activation layer to create outputs. Neurons act as a unidirectional LSTM structure in the forward state of the BLSTM. Network training can be done as a conventional unidirectional LSTM because the neurons in both networks are not linked. BLSTM designs may give better results than other network architectures depending on the scenario. For example, BLSTMs have been observed to perform effectively in voice processing applications where the content plays a vital role [15].

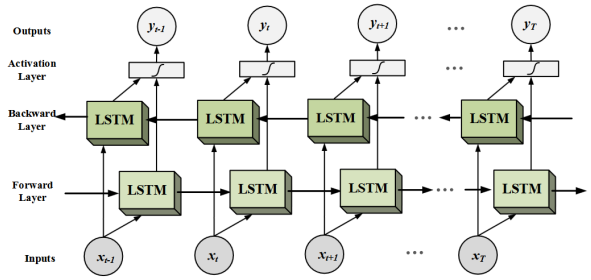

Fig 2. BLSTM design model

1.3 Feature Extraction

Feature extraction and speech recognition are two significant modules in speech recognition systems. The primary goal of feature extraction is to uncover strong and distinguishable characteristics in audio data. The recognition module decodes the spoken input using speech characteristics and acoustic models and delivers written outputs with excellent accuracy. Several approaches for extracting speech features have been endeavored so far, which include powerful features like linear predictive cepstral coefficients (LPCCs), Mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive coefficients (PLPs) .

The audio waveform itself is analyzed in temporal analysis. In spectral analysis, a speech signal's spectrum representation is examined. The main goal of feature extraction is to generate and save the feature vector sequence that best represents the input signal. First and foremost, multiple speakers record varied voice samples of each term in the corpus. Following the collection of speech samples, they are transformed from analog to digital form by sampling at a frequency of 8000Hz. Sampling is the

process of capturing voice samples at regular intervals. The obtained data is normalized in order to remove noise from speech samples if felt necessary. Following the acquisition of speech samples, feature extraction, feature training and feature testing are performed on them. The feature extraction process transforms the input signal into an internal representation that may be re-used to reconstruct the original signal. Many strategies for extracting features include MFCC, PLP, RAST, LPCC, PCA, LDA, Wavelet and DTW. MFCC is one of the mostly employed features among all these [16].

MFCC provides acoustic analysis, which depicts the ear model and demonstrates excellent results in detecting and identifying speech and speakers. Because of the capability of the ear model, MFCC is utilized to extract features, as it works in segregated mode. We can immediately distinguish a spoken word uttered by any unknown person by employing this technique. The features MFCC, ΔMFCC and ΔΔMFCC are considered together to find out the attributes of each spoken word, and an attempt has been made to find out the word recognition or detection rate. The 39 features are generated by combining 13 basic MFCC features, 13 ΔMFCC features and another 13 ΔΔMFCC features. The steps involved in extracting MFCC features are depicted in Figure 3 [17].

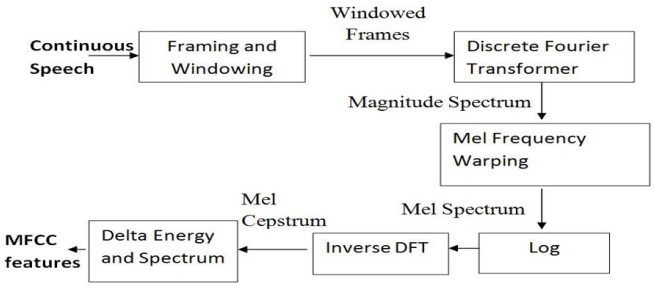

Fig 3. Feature Extraction Model

1.4 Performance Measurement

1.4.1 Classification Accuracy

An accuracy is a number that shows how well the model works for all classes in general. It's useful when all the categories are of the same importance. It is found by dividing the number of correct predictions by the total number of predictions that could be made. Accuracy can quickly calculate by confusion matrix with the help of the following formula –

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

1.4.2 Precision

The precision is the ratio of correctly classified Positive samples to all Positive samples (either correctly or incorrectly). The precision score represents how well the model can identify positive samples.

$$Precision = \frac{TP}{TP+FP}$$

1.4.3 Recall

The recall is the proportion of correctly classified positive samples to positive class samples. The recall metric assesses the model's ability to detect positives. The higher the recall, the more positive samples are discovered. The recall only concerns the classification of positive samples.

$$Recall = \frac{TP}{TP+FN}$$

1.4.4 F1 Score

Using this score, the harmonic mean of precision and recall will be calculated. To get the F1 score, a weighted average of accuracy and recall is employed. The maximum value of an F1 score is 1 and the minimum value is 0. Equal proportions of accuracy and

recall contribute to the F1 score.

$$F1\ Score = \frac{2 * Precision * Recall}{Precision * Recall}$$

2 Methodology

Speech Database

2.1 Training Data

The speech materials have been gathered by the "Indian Language Technology Proliferation and Development Center (ILTP-DC)". The dataset is 1.9 GB in size. The speech corpus comprises 32,335 spoken words by 1000 speakers, both men and women, with 262 distinct words. The gender and word information are shown in Table 1. Due to the high dB and noise, some words are deleted. The downloaded speech has a sampling frequency of 8 kHz and the speech length is in the range of 1-2 seconds.

Table 1. Gender and word information of proposed work

	Male	Female
Speech files	24,529	7806
Words	260	251

A total of 10 district names are considered for recording in this keyword spotting model as part of the data analysis. The selected utterances are contributed by various native Assamese speakers, both male and female, details of which are mentioned in Figure 4.

Sl. No	Local language	Words in English	Phone	Male	Female	Total
1.	বৰপেটা	Barpeta	BO-R-PE-TA	142	35	177
2.	বঙাইগাঁও	Bongaigaon	BO-NGA-I-GAO	120	22	142
3.	ডিব্রুগড়	Dibrugarh	DIB-RU-GO-R	338	117	455
4.	যোৰহাট	Jorhat	JOR-HA-T	179	55	234
5.	কামৰূপ	Kamrup	KA-M-RU-P	239	67	306
6.	মৰিগাঁও	Marigaon	MO-RI-GA-U	199	46	245
7.	নগাঁও	Nagaon	NO-GA-U	225	66	291
8.	নলবাৰী	Nalbari	NO-L-BA-RI	173	39	212
9.	শিৰসাগৰ	Sivasagar	SI-V-SA-GO-R	272	83	355
10.	তিনিচুকীয়া	Tinsukia	TI-NI-CU-KI-AA	159	41	200

Fig 4. Gender and phone information of spoken words

2.2 Validation Data

For validation purposes, 10% of spoken data have been separated from each district.

2.3 Proposed Model

The proposed network has six layers, as shown in Figure 5. In the first layer sequential layer, the input sequence is 39. In both BLSTM layers, the hidden units total 150. A block diagram of the proposed model is given below in Figure 6.

3 Results and Discussion

A few interesting inferences are drawn, presented here in the form of tables and figures. A few observations seem to be novel in this area of research and initiate some scope for further experiments.

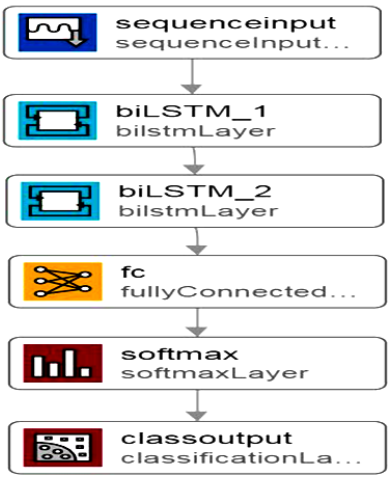

Fig 5. Layered architecture of the proposed BLSTM model

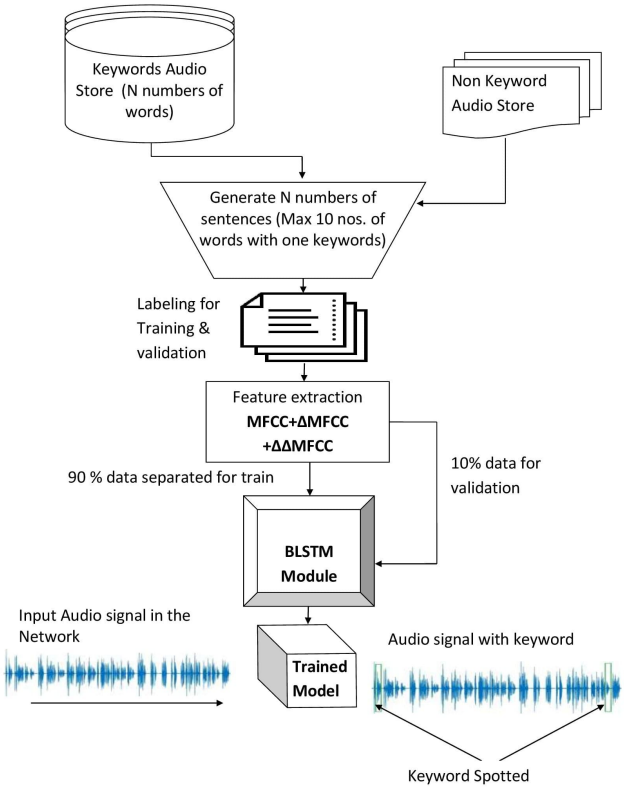

Fig 6. Block Diagram of proposed Model (BLSTM)

The word error rate of all selected spoken words are tabulated here with respect to three different features, viz., MFCC, ΔMFCC and ΔΔMFCC. But it is clearly observed in Figure 7 that the WER gets lower in percentage when we consider the ΔMFCC and ΔΔMFCC features with basic MFCC. So, it is decided to consider the MFCC, ΔMFCC and ΔΔMFCC features for further experiments in the proposed model. The space complexity is ignored for the achieving low word error rate.

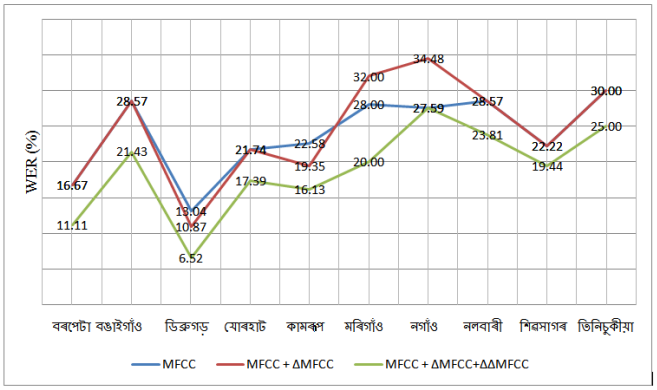

Fig 7. WER (%) graph of the selected spoken words

The proposed BLSTM model is tested in terms of WER(%), Accuracy(%), Recall, Precision and F1 score and the respective average results are found as 98.12%, 0.86, 0.95 and 0.90 for all ten words, details of which are presented through Figure 8. These results stand in favor of the model when compared to other similar results reported for other contemporary models. As mentioned earlier, the feature(s) MFCC + ΔMFCC + ΔΔMFCC, which comprises 39 coefficients, is considered to test various performance parameters of the proposed BLSTM model.

Words	WER(%)	Accuracy (%)	Recall	Precision	F1 Score
বৰপেটা	11.11	98.89	0.89	1.00	0.94
বঙাইগাঁও	21.43	97.86	0.86	0.92	0.89
ডিব্রুগড়	06.52	99.35	0.93	1.00	0.97
যোৰহাট	17.39	98.26	0.91	0.91	0.91
কামৰূপ	16.13	98.39	0.90	0.93	0.92
মৰিগাঁও	20.00	98.00	0.80	1.00	0.89
নগাঁও	27.59	97.24	0.76	0.96	0.85
নলবাৰী	23.81	97.62	0.81	0.94	0.87
শিৱসাগৰ	19.44	98.06	0.86	0.94	0.90
তিনিচুকীয়া	25.00	97.50	0.90	0.86	0.88
Average	18.84	98.12	0.86	0.95	0.90

Fig 8. WER, Accurary, Recall, Precision and F1 score of selected words

Table 2. Comparative performance analysis of other established models with the proposed model considering the dataset of the proposed model

MODEL	WER (%)	ACCURACY(%)	RECALL	PRECISION	F1 SCORE
TC-ResNet14-1.15 [10]	19.79	98.02	0.86	0.94	0.90
SincConv + DSConv [11]	20.30	97.97	0.85	0.94	0.89
Proposed Model	18.84	98.12	0.86	0.95	0.90

It is observed from Table 2 that the proposed model seems to be a promising one in terms of WER, Accuracy, Recall, Precision as well as F1 score when compared with the other two existing models, i.e., **TC-ResNet14-1.15** and **SincConv + DSConv**. It is to be noted that the same dataset as well as same features (i.e. MFCC + ΔMFCC + ΔΔMFCC) are considered throughout the experiment for all three models. It is further noticed that the two models i.e. **TC-ResNet14-1.15** and **SincConv + DSConv**

performed slightly less with their own datasets than the currently generated results with new dataset. Finally it is observed that the proposed model is performing well in all 5 measuring parameters.

4 Conclusion and Future Work

This work provides a different approach to detect specific keywords of the Assamese language by adopting a deep learning methodology. An average of 18.84% word error rate (WER) is reported along with other performance measuring parameters, viz., Accuracy, Precision, Recall and F1 Score, which are all considered promising when compared to other contemporary models. A single dataset with 39 common feature set is used throughout the entire comparative analysis. As an initial level approach, 10 native Assamese language spoken words are considered, which is subject to increase in number during further level of experiments. The combination of 39 MFCC features gives the best result for the selected model. There is a probability of reducing the WER by combining two or more network models with additional features, considering time and space complexity of the models separately, which could be tested in future approaches.

Acknowledgment

The authors would like to thank the Ministry of Electronics and Information Technology (MeitY), Government of India for consolidating and making available the linguistic resources and tools under the TDIL project through the "Indian Language Technology Proliferation and Development Center (ILTP-DC)" via http://tdil-dc.in/.

References

1) Kalita D, Borbora K. Keyword Detection using Auto Associative Neural Network with Reference to Assamese Language. *International Journal of Recent Technology and Engineering*. 2019;8(3):3290–3294. Available from: https://doi.org/10.35940/ijrte.C5428.098319.
2) Nath D, Kalita SK. A study of Spoken Word Recognition using Unsupervised Learning with reference to Assamese Language. *2019 2nd International Conference on Innovations in Electronics, Signal Processing and Communication (IESC)*. 2019;p. 98–103. Available from: https://doi.org/10.1109/IESPC.2019.8902439.
3) Lin J, Yumei Y, Maosheng Z, Defeng C, Chao W, Tonghan W. A Multiscale Chaotic Feature Extraction Method for Speaker Recognition. *Complexity*. 2020;2020:1–9. Available from: https://doi.org/10.1155/2020/8810901.
4) Georgescu ALL, Pappalardo A, Cucu H, Blott M. Performance vs. hardware requirements in state-of-the-art automatic speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*. 2021;2021(1):28–28. Available from: https://doi.org/10.1186/s13636-021-00217-4.
5) Shashidhar R, Patilkulkarni S, Puneeth SB. Combining audio and visual speech recognition using LSTM and deep convolutional neural network. *International Journal of Information Technology*. 2022;p. 1–2. Available from: https://doi.org/10.1007/s41870-022-00907-y.
6) Mahalingam H, Rajakumar MP. Speech Recognition using Multiscale Scattering of Audio Signals and Long Short-Term Memory of Neural Networks". *International Journal of Advances in Computer Science and Cloud Computing (IJACSCC)*. 2019;7(2):12–16. Available from: http://iraj.doionline.org/dx/IJACSCC-IRAJ-DOIONLINE-16658.
7) Singh A, Kaur N, Kukreja V, Kadyan V, Kumar M. Computational intelligence in processing of speech acoustics: a survey. *Complex & Intelligent Systems*. 2022;8(3):2623–2661. Available from: https://doi.org/10.1007/s40747-022-00665-1.
8) Wiesner M, Raj D, Khudanpur S. Injecting Text and Cross-Lingual Supervision in Few-Shot Learning from Self-Supervised Models. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022. Available from: arXivpreprintarXiv:2110.04863.
9) Tang R, Lin J. Deep Residual Learning for Small-Footprint Keyword Spotting. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018;p. 5484–5488. Available from: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=8450881.
10) Choi S, Seo S, Shin B, Byun H, Kersner M, Kim B, et al. Temporal Convolution for Real-Time Keyword Spotting on Mobile Devices. *Interspeech*. 2019. Available from: arXivpreprintarXiv:1904.03814,2019.
11) Mittermaier S, Kurzinger L, Waschneck B, Rigoll G. Small-Footprint Keyword Spotting on Raw Audio Data with Sinc-Convolutions. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020. Available from: arXivpreprintarXiv:1911.02086,2019.
12) Mo T, Yu Y, Salameh M, & Niu D, Jui S. Neural Architecture Search for Keyword Spotting. *Interspeech 2020*. 1982. Available from: https://doi.org/10.21437/Interspeech.2020-3132.
13) Supriya K. Trigger Word Recognition using LSTM. *International Journal of Engineering Research*. 2020. Available from: https://doi.org/10.17577/IJERTV9IS060092.
14) Araya M, Alehegn M. Text to Speech Synthesizer for Tigrigna Linguistic using Concatenative Based approach with LSTM model. *Indian Journal of Science and Technology*. 2022;15(1):19–27. Available from: https://doi.org/10.17485/IJST/v15i1.1935.
15) Sayda E, Tan KL. Speed Prediction on Real-life Traffic Data: Deep Stacked Residual Neural Network and Bidirectional LSTM". In: MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous '20). Association for Computing Machinery. ;p. 435–443. Available from: https://doi.org/10.1145/3448891.3448892.
16) Baroi OL, Kabir MSA, Niaz A, Islam MJ, Rahimi MJ. Effects of Filter Numbers and Sampling Frequencies on the Performance of MFCC and PLP based Bangla Isolated Word Recognition System. *International Journal of Image, Graphics and Signal Processing*. 2019. Available from: https://doi.org/10.5815/ijigsp.2019.11.05.
17) Yu J, Ye N, Du X, Han L. Automated English Speech Recognition Using Dimensionality Reduction with Deep Learning Approach. *Wireless Communications and Mobile Computing*. 2022;2022:1–11. Available from: https://doi.org/10.1155/2022/3597347.