

# Robust spoken word detection in assamese language using BiLSTM with data augmentation for noisy environments

Deepjyoti Kalita<sup>1</sup>, Khurshid Alam Borbora<sup>2</sup>

<sup>1</sup>Department of Computer Science and IT, Mangaldai College, Mangaldai, Assam, India

<sup>2</sup>Gauhati University Centre for Distance and Online Education, Gauhati University, Assam, India

## Article Info

### Article history:

Received Nov 21, 2024

Revised Apr 7, 2025

Accepted Jul 3, 2025

### Keywords:

BiLSTM

Data augmentation

Deep learning

Keyword detection

Machine learning

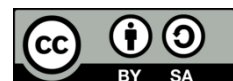
MFCC

WER

## ABSTRACT

This study focuses on enhancing spoken word detection in the Assamese language using bidirectional long short term memory (BiLSTM). The primary objective is to improve the model's robustness in noisy environments by using various data augmentation methods. The research addresses the challenges of keyword detection in low-resource languages like Assamese. A BiLSTM model was trained and tested using a speech corpus sourced from the Indian Language Technology Proliferation and Development Center (ILTP-DC), comprising 32,335 utterances from 1,000 speakers and 262 unique Assamese words. The model was trained on 10 specific keywords. Feature extraction was conducted using 39 coefficients, including MFCC,  $\Delta$ MFCC, and  $\Delta\Delta$ MFCC. The model's performance was evaluated on clean and augmented noisy datasets. The application of data augmentation techniques significantly improved the model's performance in noisy environments. This model achieved an average accuracy of 98.01% and a word error rate (WER) of 19.94% on noisy data, showcasing the effectiveness of augmentation in enhancing keyword detection. This work introduces a novel approach to Assamese spoken word detection by integrating BiLSTM with data augmentation techniques, making the model more noise-resilient. This study sets a benchmark for Assamese speech recognition and showcases augmentation techniques' effectiveness in low-resource languages.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Deepjyoti Kalita

Department of Computer Science and IT, Mangaldai College

Mangaldai, Assam

Email: deepjyoti111@gmail.com

## 1. INTRODUCTION

Spoken word detection, a crucial aspect of automatic speech recognition (ASR), involves identifying and transcribing spoken words into text. While significant advancements have been made in ASR systems for widely spoken languages, such as English and Mandarin, the development of robust speech recognition systems for low-resource languages remains a challenge. Assamese, a language spoken by millions in the northeastern region of India, represents one such low-resource language where ASR technologies are underdeveloped. The primary challenge lies in the limited availability of annotated speech corpora and the complexities of handling noise in real-world environments. In such cases, traditional methods often fail to perform at optimal levels, particularly when the model is exposed to noisy speech data [1]-[2].

The use of deep learning techniques, specifically bidirectional long short-term memory (BiLSTM), has gained popularity in the field of speech recognition. BiLSTM networks, a variation of recurrent neural networks (RNNs), are known for their ability to capture both past and future context in sequential data,

making them ideal for speech recognition tasks. Unlike traditional unidirectional long short-term memory (LSTMs), BiLSTM processes data in both forward and backward directions, enhancing its ability to recognize spoken words in a more accurate and context-aware manner. The effectiveness of BiLSTM in spoken language processing has been demonstrated in several studies, such as [1]-[2], which showed improvements in speech recognition accuracy by leveraging BiLSTM's ability to capture temporal dependencies in audio signals.

However, a persistent challenge in ASR systems, especially for low-resource languages, is the performance degradation caused by noisy data. Environmental noise, such as background chatter, reverberations, and distortions, can significantly reduce the accuracy of speech recognition systems. To combat this, data augmentation techniques have been explored as a solution to enhance the robustness of speech recognition models. Data augmentation methods, such as noise injection, time-stretching, and pitch-shifting, simulate real-world conditions and provide the model with diverse training data to improve generalization. Studies like [3] and [4] have shown that data augmentation can substantially improve the performance of ASR systems by making them more resilient to noise and improving their ability to detect keywords in varied acoustic conditions [5]-[10].

In the context of Assamese, the need for effective keyword detection systems is even more pronounced due to the scarcity of publicly available speech corpora and the linguistic diversity within the language. Assamese, a language rich in phonetic diversity, has posed a significant challenge for ASR development. Although some efforts have been made to develop ASR systems for Assamese, they have been limited by the lack of large annotated datasets and linguistic resources. The Assamese speech corpus provided by the Indian Language Technology Proliferation and Development Center (ILTP-DC) is one of the few available resources, yet it still contains limitations regarding the diversity of speech patterns, background noise, and regional variations. Recent researches has focused on building rudimentary ASR systems for Assamese, but these approaches have yet to fully address issues such as noise resilience and real-time keyword detection [11]-[14].

Data augmentation in the context of Assamese speech recognition is still an emerging area. Techniques like noise injection, pitch scaling, and reverberation simulation have shown promise in improving performance, especially in noisy environments. Augmenting Assamese speech data has the potential to boost recognition accuracy, particularly when dealing with low-resource settings and noisy acoustic conditions. This paper aims to contribute to this research by utilizing BiLSTM in combination with data augmentation methods to enhance spoken word detection in Assamese, providing a more noise-resilient and effective solution for this underrepresented language [15], [16].

## **2. METHOD**

### **2.1. Standard speech corpus**

The ILTP-DC has created a comprehensive corpus for the Assamese language that includes both text and speech data. This corpus is an important resource for researchers working on natural language processing (NLP) and speech-related technologies. It contains various types of written content such as news articles, books, and documents, alongside a large speech dataset. The speech data is particularly useful for developing speech recognition and other spoken language processing technologies.

The ILTPDC corpus for Assamese includes 32,335 spoken words recorded by 1,000 speakers, both male and female. It covers 262 distinct words, with the audio sampled at 8 kHz and a length of 1-2 seconds per word. Due to noise and high dB levels, some words were removed from the corpus. The dataset includes 24,529 male speech files and 7,806 female speech files, with the number of words recorded being 260 for male speakers and 251 for female speakers. This corpus is crucial for advancing speech recognition technologies in the Assamese language [2]. A total of 10 district names in Assam are used for recording in the keyword spotting model, with data collected from a variety of male and female native Assamese speakers.

### **2.2. Preprocessing**

For preprocessing, we enhance our dataset by applying data augmentation methods, which help expand the dataset in terms of both size and diversity. By modifying aspects of the speech, such as pitch, speed, volume, and adding noise, we generate new variations of the data. This technique not only increases the amount of training data but also improves the robustness and generalization of the model. With this enriched dataset, the model becomes better equipped to handle variations in accents, speech rates, and environmental noise, leading to improved accuracy and performance in speech recognition tasks. This approach has been widely adopted in speech recognition systems to improve recognition accuracy, particularly in low-resource settings and noisy environments.

### 2.3. Data augmentation

Data augmentation is a technique used to expand and diversify the speech dataset in a controlled manner. It involves applying various modifications to the original speech signals, such as altering pitch, speed, and volume, as well as adding or removing background noise. The primary goal of data augmentation is to enhance the robustness and generalization of models. By training models on larger, more varied datasets, they can better handle differences in accent, speaking rate, and background noise, leading to improved accuracy. Two augmentation methods have been applied in my experiment.

- Time stretching: this time stretching technique is applied to change the duration of a speech signal where the pitch remains the same. The time-domain processing of audio signals like the algorithms of time-scale modification is applied for this purpose. It has also been used to simulate faster or slower variations in the speaking rate. For training ASR models for recognizing speeches by different people in different situations at different speeds, this technique is effectively applied.
- Adding noise: adding noise involves adding background noise to a speech signal subject to simulate recording in different conditions, adding noise is involved. It is very effective in a noisy environment or a distant microphone. For training ASR models in recognizing speech in noisy environments, it is very useful. Many different types of noises namely white noise, pink noise and babble noise are added to a speech signal as per the requirements [17].

### 2.4. Feature extraction

Feature extraction and speech recognition are essential components of speech recognition systems. The primary aim of feature extraction is to capture the most relevant and distinguishing characteristics from the audio signal. On the other hand, the recognition module uses these features along with acoustic models to convert spoken words into written text with high accuracy.

Various techniques have been used for extracting features from speech, including linear predictive cepstral coefficients (LPCCs), mel-frequency cepstral coefficients (MFCCs), and perceptual linear predictive coefficients (PLPs). These features are derived through two main types of analysis: temporal analysis, which focuses on the changes in the audio signal over time, and spectral analysis, which examines the frequency components of the signal.

The process begins by recording voice samples from multiple speakers. These samples are then digitized by sampling them at a frequency of 8,000 Hz. After this, any noise in the data may be removed through normalization. The feature extraction process then transforms the raw audio data into a more useful form, making it easier for the recognition system to process.

Several methods can be used for feature extraction, including MFCC, PLP, LPCC, and others like principal component analysis (PCA), linear discriminant analysis (LDA), wavelet transform, and dynamic time warping (DTW). Among these, MFCC is one of the most commonly used due to its excellent performance in modeling how humans hear speech. It is particularly effective in recognizing speech and identifying speakers.

MFCC, along with its derivatives  $\Delta$ MFCC and  $\Delta\Delta$ MFCC, are combined to generate a feature vector. This combination captures not just the static properties of the speech but also its dynamics, providing a comprehensive representation of each spoken word. The 39 features are created by combining the 13 basic MFCC features, 13  $\Delta$ MFCC features, and 13  $\Delta\Delta$ MFCC features. This method of feature extraction allows for improved recognition of spoken words, even by unknown speakers, and is widely used in many speech recognition systems [18]-[20].

### 2.5. Mechanism of Bi-LSTM:

The BiLSTM model is an extension of the standard LSTM architecture, designed to improve the processing of sequential data by incorporating context from both past and future inputs. This makes it especially powerful for applications such as speech recognition, text analysis, and time-series prediction [21]-[24]. The mechanism how BiLSTM works discussed.

- Bidirectional processing: unlike traditional LSTMs, which process data in a unidirectional manner (i.e., from past to future), a BiLSTM processes data in two directions: forward and backward. The forward LSTM reads the input sequence from the beginning to the end, while the backward LSTM reads the sequence in reverse, from end to beginning. This dual flow allows the model to capture both past and future context, which can be particularly beneficial in tasks like speech recognition and NLP where future context (future words or sounds) can be as important as past context.
- Memory cells: at the core of LSTMs is the memory cell, which helps in storing information for long durations. This is achieved through gates-input, forget, and output-that control the flow of information. The memory cell ensures that important information is retained over longer sequences, addressing the vanishing gradient problem faced by traditional RNNs in long-term sequence learning. In BiLSTMs,

both the forward and backward LSTMs have independent memory cells, allowing them to learn complementary features from the two directions.

- Gates and update mechanism: the input gate decides which values to update, the forget gate determines which values to discard, and the output gate controls which values are outputted. These gates are regulated by learned weights, which are updated through backpropagation during the training process. The gates in the forward and backward LSTMs are updated independently but follow the same mechanism.
- Training and applications: during training, the BiLSTM is fed with sequences, and the model learns to adjust its weights to minimize a loss function (such as cross-entropy loss for classification tasks). This result in a model that can predict future states based on both past and future context. BiLSTMs have been used in various domains, including speech recognition (e.g., recognizing spoken words in noisy environments), language translation, sentiment analysis, and named entity recognition in NLP tasks.

By processing data in both directions, BiLSTMs can leverage information from both the future and the past, which is essential for tasks like speech-to-text conversion, where the full context of a word can be necessary to understand it correctly. Because BiLSTMs take into account both future and past data, they are less dependent on sequence length compared to simpler models [25]-[26]. The basic operational model of a BiLSTM network is illustrated in Figure 1.

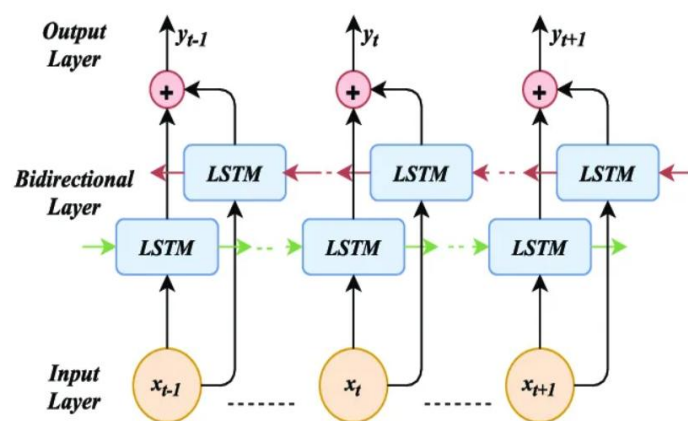


Figure 1. Bi-LSTM model design [21]

## 2.6. Performance measurement

Performance measurement is essential to evaluate how well the BiLSTM model performs in keyword recognition. It allows us to assess the correctness of predictions and the model's ability to detect keywords accurately. For this purpose, four widely used metrics are considered: accuracy, precision, recall, and F1 score.

- Accuracy: classification accuracy measures how well the model performs overall. It tells us the proportion of correct predictions out of all the predictions made. This metric is especially useful when all classes are equally important. A higher accuracy indicates that the model is making more correct predictions across all categories.
- Precision: precision focuses on the model's ability to correctly identify positive instances. It measures how many of the predicted positive results are actually correct. A high precision means that the model is good at avoiding false positives, or in other words, it doesn't frequently predict positives when they are actually negatives.
- Recall: recall measures how well the model detects positive samples. It looks at how many of the actual positives are correctly identified by the model. A higher recall indicates that the model is good at detecting most of the positive cases, reducing the chances of missing important instances.
- F1-score: the F1-score is a balance between precision and recall. It is used when there is a need to consider both the accuracy of positive predictions (precision) and the model's ability to detect all relevant positive samples (recall). The F1 score is particularly useful when the data is imbalanced, and it provides a single metric that captures both aspects. Higher F1-scores indicate a better balance between precision and recall [27], [28].

## 2.7. Proposed algorithm

The following Algorithm 1 shows the steps to build and train a BiLSTM model for keyword recognition in speech.

### Algorithm 1. Keyword recognition using BLSTM

**Input:** Recorded keyword, Non-keyword audio store

**Output:** Trained BLSTM model for keyword recognition

1. Store the recorded keyword in an audio store;
2. Increase the number of recorded keywords using data augmentation to create a larger dataset (N words);
3. Store non-keywords in another audio store (N words);
4. Generate N synthetic sentences (1 keyword with 10 non-keywords);
5. Normalize the dataset for smoothness;
6. Label the dataset for training & validation data;
7. Divide the dataset into a 90:10 ratio for training and validation;
8. Extract features of speech signals (varies in each experiment, e.g., MFCC,  $\Delta$ MFCC,  $\Delta\Delta$ MFCC, etc.);
9. Design a BLSTM model with 6 layers:
  - o One sequence input layer
  - o Two Bi-LSTM layers
  - o One fully connected layer
  - o One softmax layer
  - o One classification layer
10. Provide input to the BLSTM model for training and validation;
11. Repeat the training process until an acceptable accuracy is achieved;
12. For testing, feed random sentences labeled with keywords and non-keywords to the trained model; 13. Model returns values for each frame of test signals. If values are high, mark as Keyword; otherwise, mark as Non-keyword;

## 3. RESULTS AND DISCUSSION

The experiments were conducted under three different conditions to evaluate the system's performance. Initially, the model was tested on normalized, noise-free data. Consequently, additional noise was introduced into the dataset, and the experiments were repeated to assess the impact of noise on recognition performance. Finally, data augmentation techniques were applied to expand the dataset, and experiments were conducted to determine the improvements achieved. The results of these experiments are summarized in Table 1 and discussed in detail below.

Table 1. Comparison of result of our experiment

Models (BiLSTM) with different dataset	WER (%)	Accuracy (%)	Recall	Precision	F1-score
Normalized data	18.84	98.12	0.86	0.95	0.90
Adding extra noise	23.90	97.61	0.84	0.91	0.88
With data augmentation and extra noise	<b>19.94</b>	<b>98.01</b>	<b>0.88</b>	<b>0.92</b>	<b>0.90</b>

### 3.1. Analysis using noise factor

Under normal conditions, our model performs well with a word error rate (WER) of 18.84%, indicating accurate word recognition. It achieves an accuracy of 98.00%, recall of 0.86, precision of 0.95, and an F1 Score of 0.90, displaying robust ability to minimize false positives. However, it is observed that the model's performance is influenced when extra noise is added to the dataset. Though the WER increases to 23.90%, there are slight reductions in recall, precision, and F1-score. This implies clearly that the model is sensitive to the presence of additional noise, leading to a decline in overall recognition accuracy.

### 3.2. Analysis using noise factor after data augmentation

From the result mentioning in Table 1, it can be clearly observed before using data augmentation, this model made mistakes in understanding words i.e., WER about 23.90% of the time. But after adding some variations through data augmentation, the mistakes reduced to 19.94%. It's like the model got better at getting the words right. The accuracy of the model also went up slightly from 98.00% to 98.01%. For other measures like recall, precision, and F1-score, it can see improvements too. These measures help understand how well the model is doing. The decrease in mistakes WER suggests that the model got more robust,

meaning it became better at handling different ways people speak. Overall, data augmentation helped the model balance its accuracy, making it more effective.

#### 4. CONCLUSION

This study shows that BiLSTM networks are effective for recognizing spoken words in Assamese. The model can understand words better by considering both past and future sounds. Our results highlight the importance of reducing noise and using data augmentation to make the model more accurate in real-world situations. The BiLSTM model handled noisy environments well, and data augmentation further improved its performance. This makes it a strong choice for speech recognition in low-resource languages like Assamese.

For future work, the model can be improved by adding more layers or trying other approaches like BiLSTM-CNN or transformer models. Expanding the dataset to include different dialects, improving how the model deals with noise and using advanced techniques like transfer learning and speaker adaptation can further improve accuracy. These steps will help develop better speech recognition systems for Assamese and other low-resource languages. This study shows that BiLSTM models can recognize Assamese spoken words well, but noise affects accuracy. However, data augmentation helps the model perform better. To improve speech recognition for Assamese, future research should expand datasets, explore new model designs, and test in real-world applications.

#### ACKNOWLEDGMENTS

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/>.

#### FUNDING INFORMATION

Authors state no funding involved.

#### AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Deepjyoti Kalita	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Khurshid Alam		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓		
Borbora														

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

#### CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

#### DATA AVAILABILITY

The data that support the findings of this study are available from Indian Language Technology Proliferation and Development Center (ILTP-DC). Restrictions apply to the availability of these data, which were used under license for this study. Data are available [from the <http://tdil-dc.in/>] with the permission of ILTP-DC.






## REFERENCES




- [1] X. Li *et al.*, "Bidirectional LSTM network with ordered neurons for speech enhancement," in *Interspeech 2020*, ISCA: ISCA, Oct. 2020, pp. 2702–2706. doi: 10.21437/Interspeech.2020-2245.
- [2] D. Kalita, K. A. Borbora, and D. Nath, "Use of bidirectional long short term memory in spoken word detection with reference to the assamese language," *Indian Journal Of Science And Technology*, vol. 15, no. 27, pp. 1364–1371, Jul. 2022, doi: 10.17485/IJST/v15i27.655.
- [3] M. Dua, Akanksha, and S. Dua, "Noise robust automatic speech recognition: review and analysis," *International Journal of Speech Technology*, vol. 26, no. 2, pp. 475–519, Jul. 2023, doi: 10.1007/s10772-023-10033-0.
- [4] M. Huh, R. Ray, and C. Karnei, "A comparison of speech data augmentation methods using S3PRL toolkit," Feb. 2023, [Online]. Available: <http://arxiv.org/abs/2303.00510>
- [5] J. Lin, Y. Yumei, Z. Maosheng, C. Defeng, W. Chao, and W. Tonghan, "A multiscale chaotic feature extraction method for speaker recognition," *Complexity*, vol. 2020, pp. 1–9, Dec. 2020, doi: 10.1155/2020/8810901.
- [6] A.-L. Georgescu, A. Pappalardo, H. Cucu, and M. Blott, "Performance vs. hardware requirements in state-of-the-art automatic speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, p. 28, Dec. 2021, doi: 10.1186/s13636-021-00217-4.
- [7] R. Shashidhar, S. Patilkulkarni, and S. B. Puneeth, "Combining audio and visual speech recognition using LSTM and deep convolutional neural network," *International Journal of Information Technology*, vol. 14, no. 7, pp. 3425–3436, Dec. 2022, doi: 10.1007/s41870-022-00907-y.
- [8] H. Mahalingam and M. P. Rajakumar, "speech recognition using multiscale scattering of audio signals and long short-term memory of neural networks," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 11, pp. 2955–2962, 2019, doi: 10.35940/ijitee.K2270.0981119.
- [9] A. Singh, N. Kaur, V. Kukreja, V. Kadyan, and M. Kumar, "Computational intelligence in processing of speech acoustics: a survey," *Complex & Intelligent Systems*, vol. 8, no. 3, pp. 2623–2661, Jun. 2022, doi: 10.1007/s40747-022-00665-1.
- [10] M. Wiesner, D. Raj and S. Khudanpur, "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)*, Singapore, Singapore, 2022, pp. 8597–8601, doi: 10.1109/ICASSP43922.2022.9746852.
- [11] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 5484–5488, doi: 10.1109/ICASSP.2018.8462688.
- [12] S. Choi *et al.*, "Temporal Convolution for Real-time Keyword Spotting on Mobile Devices," Apr. 2019, [Online]. Available: <http://arxiv.org/abs/1904.03814>.
- [13] B. Deka, S. R. Nirmala, and S. K., "Development of assamese continuous speech recognition system," in *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, ISCA: ISCA, Aug. 2018, pp. 220–224. doi: 10.21437/SLTU.2018-46.
- [14] D. Kalita and K. A. Borbora, "Keyword detection using auto associative neural network with reference to assamese language," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 3, pp. 3290–3294, Sep. 2019, doi: 10.35940/ijrte.C5428.098319.
- [15] M. K. Majhi and S. K. Saha, "An automatic speech recognition system in Odia language using attention mechanism and data augmentation," *International Journal of Speech Technology*, vol. 27, no. 3, pp. 717–728, Sep. 2024, doi: 10.1007/s10772-024-10132-6.
- [16] R. Zevallos, N. Bel, G. Cámbara, M. Farrús, and J. Luque, "Data augmentation for low-resource quechua ASR improvement," Jul. 2022, [Online]. Available: <http://arxiv.org/abs/2207.06872>
- [17] Y. Tada, Y. Hagiwara, H. Tanaka, and T. Taniguchi, "Robust understanding of robot-directed speech commands using sequence to sequence with noise injection," *Frontiers in Robotics and AI*, vol. 6, Jan. 2020, doi: 10.3389/frobt.2019.00144.
- [18] A. Brueggeman, T. Higuchi, M. Delfarah, S. Shum, and V. Garg, "Does single-channel speech enhancement improve keyword spotting accuracy? a case study," Sep. 2023, [Online]. Available: <http://arxiv.org/abs/2309.16060>.
- [19] M. Yu, X. Ji, B. Wu, D. Su, and D. Yu, "End-to-End Multi-Look Keyword Spotting," May 2020, [Online]. Available: <http://arxiv.org/abs/2005.10386>.
- [20] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," Apr. 2018, [Online]. Available: <http://arxiv.org/abs/1804.03209>.
- [21] I. K. Ihianle, A. O. Nwajana, S. H. Ebeunuwa, R. I. Otuka, K. Owa, and M. O. Orisatoki, "A deep learning approach for human activities recognition from multimodal sensing devices," *IEEE Access*, vol. 8, pp. 179028–179038, 2020, doi: 10.1109/ACCESS.2020.3027979.
- [22] S. Mittermaier, L. Kürzinger, B. Waschneck, and G. Rigoll, "Small-footprint keyword spotting on raw audio data with sinc-convolutions," Nov. 2019, [Online]. Available: <http://arxiv.org/abs/1911.02086>.
- [23] T. Mo, Y. Yu, M. Salameh, D. Niu, and S. Jui, "Neural architecture search for keyword spotting," in *Interspeech 2020*, ISCA: ISCA, Oct. 2020, pp. 1982–1986. doi: 10.21437/Interspeech.2020-3132.
- [24] Kalyanam Supriya, "Trigger word recognition using LSTM," *International Journal of Engineering Research and*, vol. V9, no. 06, Jun. 2020, doi: 10.17577/IJERTV9IS060092.
- [25] M. Araya and M. Alehegn, "Text to speech synthesizer for tigrigna linguistic using concatenative based approach with LSTM model," *Indian Journal of Science and Technology*, vol. 15, no. 1, pp. 19–27, Jan. 2022, doi: 10.17485/IJST/v15i1.1935.
- [26] S. Elmi and K.-L. Tan, "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*, New York, NY, USA: ACM, Dec. 2020, pp. 435–443. doi: 10.1145/3448891.3448892.
- [27] O. L. Baroi, Md. S. A. Kabir, A. Niaz, Md. J. Islam, and Md. J. Rahimi, "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*, vol. 11, no. 11, pp. 36–42, Nov. 2019, doi: 10.5815/ijgsp.2019.11.05.
- [28] J. Yu, N. Ye, X. Du, and L. Han, "Automated english speech recognition using dimensionality reduction with deep learning approach," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–11, Mar. 2022, doi: 10.1155/2022/3597347.

---

**BIOGRAPHIES OF AUTHORS**

**Deepjyoti Kalita**    completed his graduation (BCA) from Dispur College under Gauhati university. He did his masters (M.Sc.) from Gauhati University in Subject Computer Science. He obtained his Ph.D. From the Department of Computer Science, Gauhati University, Assam, India. He is currently working as assistant professor at Mangaldai College with teaching experience of over 11 years. His research interests are signal processing, networking, and data mining. He can be contacted at email: [deepjyoti111@gmail.com](mailto:deepjyoti111@gmail.com).



**Khurshid Alam Borbora**    is an assistant professor of Computer Science, at Gauhati University Centre for Distance and Open Education (GUCDOE), Gauhati University, Assam, India. He obtained his Ph.D. From the Department of Computer Science, Gauhati University, Assam, India. He has a teaching experience of over 15 years and is involved in areas of expert systems, biometrics, medical image processing, and speech processing. He can be contacted at email: [khurshidborbora007@yahoo.co.in](mailto:khurshidborbora007@yahoo.co.in).