Low Resource Indian Spoken Language Identification Using Bi-Directional Long Short Term Memory Network

Ungkar Saikia[a], Subhrojit Saikia[b], Monita Wahengbam[c], Pronob Jyoti Saikia[d]

[a]Assam Engineering College, Guwahati, Assam

[b,c]National Institute of Electronics and Information Technology, Jorhat Extension Center, Jorhat, Assam

[c]Jagannath Barooah College (Autonomous), Jorhat, Assam

Abstract

Language Identification (LID) is the process by which a speaker's language is identified using speech samples and machine learning algorithms. LID can be thought of as the frontend of Automatic Speech Recognition (ASR) systems. There are many challenges faced while developing an Indian LID given India is a multilingual country. There are many real-world practical benefits of an Indian LID. This paper focuses on an Indian Language Identification system that uses a custom model based on an Recurrent Neural Network-Long Short Term Memory (RNN-LSTM) architecture combined with MFCC feature extraction methods for speech detection and identification. The targeted languages for the proposed LID include Assamese, Kannada, Hindi, Malayalam, and Telugu.

Proposed LID uses a bi-directional LSTM architecture for feature learning. It is used in conjunction with a Mel Frequency Cepstral Coefficients (MFCC). MFCC focuses on the physical characteristics of an audio signal. One hot encoding is used to format the categorical values to a numerical format.

The results with the proposed system achieve good accuracy results with the numbers reaching a high of about 99.1% showcasing the effectiveness of the developed model in language classification tasks and comparing well with systems with more complex architectures.

Keywords: Language Identification, Deep Learning, Indian Languages, Bi-Directional LSTM

1. Introduction

Accurately recognizing or identifying the language of a speaker using a machine learning or deep learning on speech data is known as Language Identification (LID) [1a][2]. India being one of the most diverse countries in the world with respect to language variety, possess a unique challenge of identifying a speakers' languages in various contexts. While humans are an excellent language identifier, the variety of languages and dialects may pose a problem to even the most knowledgeable humans. This is much harder in part because India has many different language families with the two most common being Indo-Aryan and Dravidian language families. This diversity in languages make Language Identification important and prevalent in the context of India. With ample research in the field of LID systems, there have been ample opportunities for the real-life application of LIDs. These real-life applications include speech recognition, translation, multilingual communication services such as in call centers etc. LIDs [3] have various speech processing applications such as *automatic speech recognition* (ASR) [4]*, spoken emotion recognition* (SER) [5]*, or speaker recognition* (SR) [6].

Languages or speech or audio in general have various levels of abstractions. Also, language itself have many linguistic dimensions associated with it such as phonemes, prosodic, phonotactic information, syllable structure, prosodic, lexical words and grammar etc. [3] which fall in spoken level and word level characteristics of a language. A LID system takes in context all of these dimensions and abstractions to make a better, more accurate system. Also, Audio has various classifications like, high-level, mid-level, low-level, time and frequency domains etc. These abstractions and classifications often overlap each other, often falling in different categories. One such classification is the MFCC feature extraction which is a mid-level, time-frequency domain representation. MFCCs capture the spectral characteristics of the speech signal, which are crucial for differentiating between languages. MFCCs capture the spectral characteristics of the speech signal, which are crucial for differentiating between languages. The MFCCs basically convert physical features of audio data into array-based coefficients which act as feature representations. These are then feed in a deep learning model for language detection and identification. This research work focuses on a using a LSTM-RNN that captures both past and future context for better understanding of the data. The remaining paper is organized into the following 6 sections. Section 2 of the paper reviews previous work done in this field. Section 3 gives a detailed description of the methodology used in the research. Section 4 focuses on the experiments done. Section 5 presents the discussions and the resulting conclusions. Lastly, future works on the subject are discussed in Section 6.

2. Literature Survey

This section focuses on the previous work done in this research field. In past many different approaches having been taken for LID systems. We will mainly look at recent work done on deep learning LID systems for Indian languages. The recent work in language identification systems have led to advancements in this field leading to higher accuracy and enhanced efficiency, by leveraging different types of deep learning models and feature extraction methods.

Ahmed et al. [7] developed a Convolutional Neural Network (CNN) based approach LID using Mel spectrograms for feature extraction for Kashmiri, Hindi, Dogri, Ladakhi and other languages. The study utilized data from IIIT-H, Vox Forge and various local TV channels for speech data. The model achieved 100% testing accuracy with 100 epochs for training. This research was the first LID system to focus on the Ladakhi language making it unique in that regard.

Basu et al [8] addressed the challenge of SIDs and LIDs in low resource Eastern and North-Eastern Indian languages. Their system employed various hybrid feature extraction methods such as MFCC, SDC, RASTA-PLP that were combined with LSTM, GMM, VNN for training and testing. Apart from this, i-vectors, time delay neural networks (TDNN), were being experimented with to comply with the recent approaches. Various combination of the feature extraction and deep learning models were used with a self-recorded audio corpus. The best SID and LID performances were observed to be 94.49% and 95.69%, respectively, for the baseline systems using LSTM-RNN with MFCC+SDC feature. These highlighted the potential of LSTM-RNN systems and various hybrid feature extraction methods in handling low-resource scenarios.

Godbole et al. [9] proposed a method that used CNN (Convolutional Neural Network) for language classification. Their research converted the audio samples in .WAV format into spectrogram visuals. These spectrograms visualize changes in frequencies within an audio sample over time. Three spectrograms were generated namely Log Spectrogram, Gammatonegram, IIR-CQT Spectrogram using audio samples from a standardized IIT-H Indic Speech Database. The research was able to achieve an accuracy of 98.86% with the proposed methodology.

Deshwal et al. [1b] created a language identification system using hybrid features and a feed forward back-propagation neural network. Performance of the model was measured using various combinations of the feature extraction methods such as MFCC, PLP, combined with their 1st order derivatives, MFCC + RASTA-PLP, MFCC + SDC (Shifted delta cepstral coefficients) and all of them were compared. The model achieved highest classification rates of 94.6% with a minimum test error rate of 0.10 with MFCC + RASTA-PLP hybrid features utilizing “trainlm” learning function with a user defined database on Tamil, English, Malayalam and Hindi Languages. The study highlighted the benefits of using different training algorithms with different feature extraction methods to optimize performance using the same data as the baseline of comparisons.

Sisodia et al. [10] devised a method that used Mel Frequency Cepstral Co-efficient (MFCC) feature extraction method as primary features to retrieve necessary information from the samples and also used ensemble learning methods to classify spoken languages. Ensemble [11] learning models create a collection of classifiers and after that categorize new data samples by considering an independent choice of their calculations. The various ensemble classifier methods such as

Bagging, Boosting, Random Forest etc. were able to achieve accuracy scores ranging between 76% to 85.4 %.

Guha at el. [12] developed a Spoken Language Identification (S-LID) system that used feature extraction methods like Mel Spectrograms features and Relative Spectral Transform - Perceptual Linear Prediction (RASTA-PLP) features with new hybrid Feature Selection (FS) algorithm have been developed using the versatile Harmony Search (HS) algorithm and a new nature-inspired algorithm called Naked Mole-Rat (NMR) algorithm to select the best subset of features and reduce the model complexity to help it train faster. The selected feature set were fed to five state-of-the-art classifiers namely Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), Multi-layer Perceptron (MLP), Naïve Bayes (NB) and Random Forest (RF) for the S-LID system. The audio samples used for training the S-LID were taken from a collection of Single Speaker Speech Datasets for 10 Languages, CSS10, VoxForge and Indic TTS database from Indian Institute of Technology, Madras (IIT-Madras) speech corpus database. The results obtained from the research found out that RF classifier has achieved the highest classification accuracy of 99.89% using the CSS10 database, 99.82% using VoxForge database and 99.75% using Indic TTS database and an improvement for their raw feature set counterparts. The results obtained from each of these three datasets using all the classifiers support the claim that FS is an effective step in the learning process of the model

Ranga et al. [13] explored spectral augmentation techniques to improve code-switched language identification. By implementing spectral augmentation on a CNN-LSTM(Bi-directional) architecture with the addition of a CTC loss function, the study reported a relative improvement of 3-5% in LID accuracy over a baseline system proposed by Microsoft. The study underscores the importance of spectral augmentation in enhancing the performance of LID systems. Spectral augmentation helps in LID system more robust to variations in speech caused by various factors such as noise, accent, style etc.

Madhu et al. [14] developed a feed forward neural network for an automatic language identification system across seven Indian language. The proposed system used higher level language dependent phonotactic features and prosodic information. In phonotactic approach, a multilingual PE is used to obtain the phoneme sequence of the input speech utterance. In prosody-based approach, feature vectors are obtained by concatenating the features of three consecutive syllables**.** A feed forward neural network classifier is used at the back-end for obtaining the language identity of the given speech utterance. For the same set of test utterances, the accuracy of the phonotactic system is 72% and for the prosodic system, it is 68%.

3. Methodology

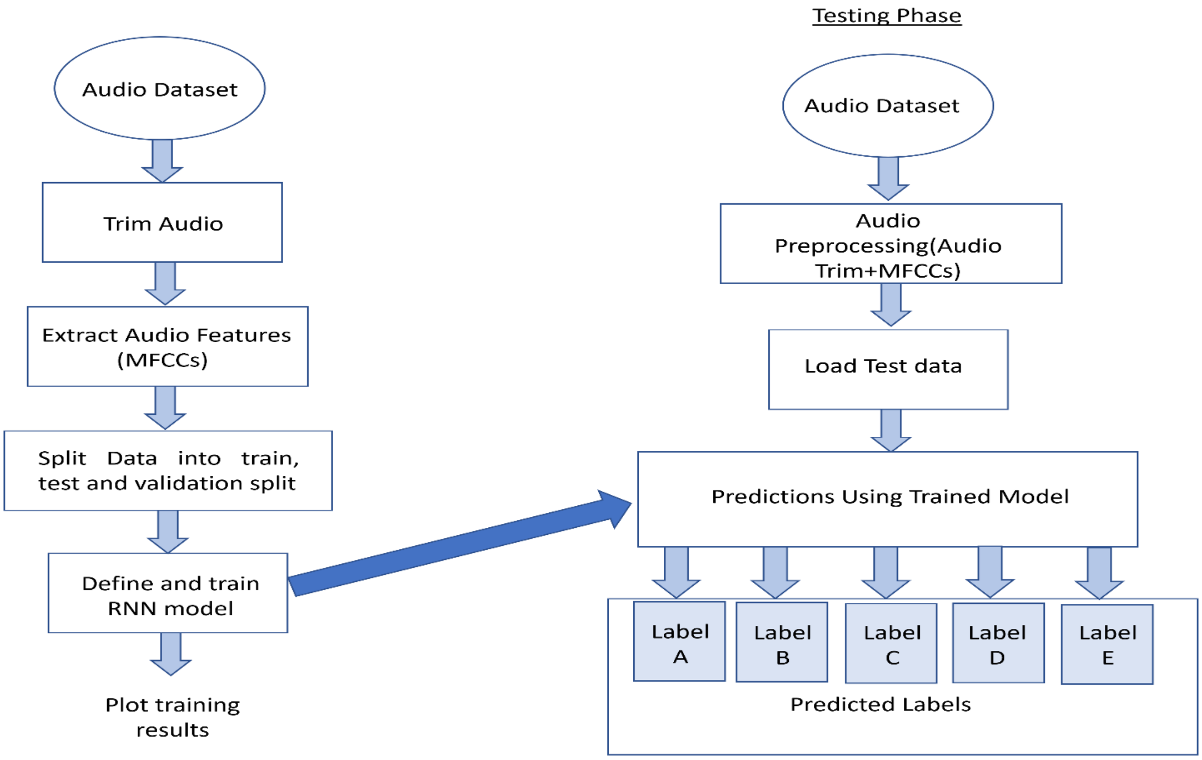

Testing Phase
Audio Dataset
Trim Audio
Extract Audio Features (MFCCs)
Split Data into train, test and validation split
Define and train RNN model
Plot training results
Audio Dataset
Audio Preprocessing(Audio Trim+MFCCs)
Load Test data
Predictions Using Trained Model
Label A
Label B
Label C
Label D
Label E
Predicted Labels

Fig 1: System Workflow

3.1 Dataset

For this research work, the data has been collected from multiple audio datasets from Kaggle and then compiled to one single dataset. This consolidated dataset contains audio samples in various languages, with each language category stored in a separate directory. The dataset includes audio files in different file formats such as MP3, WAV, etc. The total number of files present in the dataset is 5000 files. To make sure that the machine learning model properly trains on the given dataset, the audio files have to formatted and standardized to a common format and length. To improve accuracy, background noise is removed from the data. Also, padding the audio sequences with silence is done to ensure all samples have the same duration. This step is crucial for maintaining consistent input sizes for the model during training, as neural networks typically require fixed-size inputs. The audio sampling frequency is of 44 kHz with .mp3 format.

Also, using One Hot Encoding, the directory names (language labels) are encoded as numerical values using integer encoding. This ensures that the categorical values are converted to a binary representation, which helps the deep learning algorithm to handle the data better and thereby improve their performance. One hot encoding also helps in keeping the nominal relationship within the various data samples.

3.2 Feature Extraction

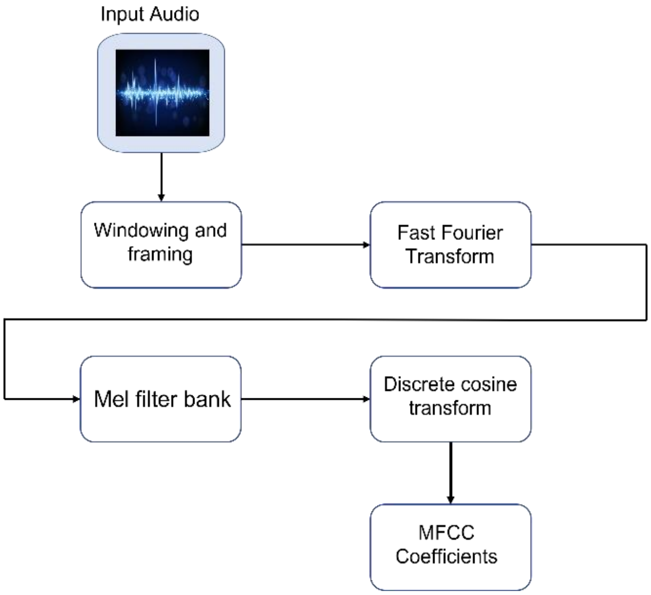


Fig 2: MFCC Feature Extraction

MFCC (Mel Frequency Cepstral Coefficients) is utilized to extract relevant features from audio samples, which are then used for training the machine learning model. MFCC acts as the input data for training of the model. MFCC basically converts the audio data features into machine readable NumPy format.

MFCC are based on human auditory response and they capture the distinctive aspects of speech and audio signals, such as spectral characteristics and timbral qualities. This enables the model to focus on relevant information for classification tasks.

The steps involved with MFCC feature extraction are:

Windowing: In this step, audios are chopped up into frames called windows which are then overlapped. This is done so as to properly focus on the audio segments and how frequency changes over time and overlapping is done to preserve the audio context. Here, window of size n-fft =2048 is used for sampling, and hops of size 512 are used for overlapping the frames.

Fourier Transform: Using a mathematical function called Discrete Fourier transform, the audios are converted from frequency domain to time domain.

Mel Filterbank: Triangular filters are used on the audio spectrum. These filters mimic human auditory perception and basically filter or compress out the frequencies that are not very perceptible to human ears. These filters are placed on a scale called the Mel scale, that mimics how human brains process pitch. The conversion from hertz to Mel is given by

$$m = 2595 \log10 (1+f/700)$$

The output of the Mel filterbanks is then computed over time to see changes in spectral characteristics with two functions called delta and delta-delta.

Discrete Cosine Transform (DCT): The outputs are then further compressed and then converted back to the time domain using and inverse FFT called Discrete Cosine Transform. The resulting outputs are the MFCCs.

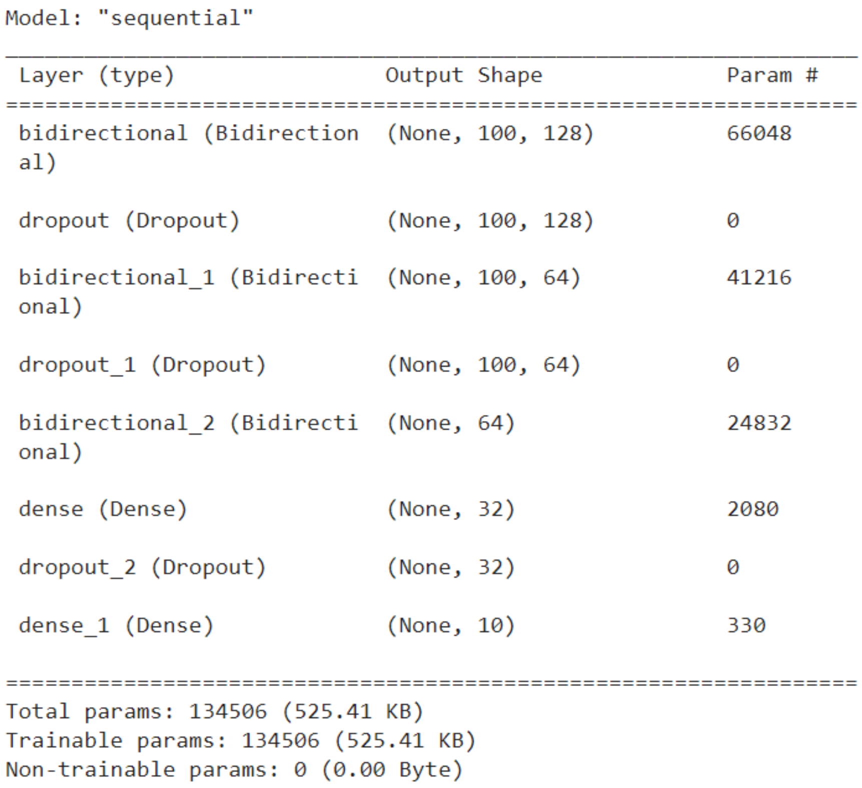
After computing the MFCC coefficients, normalization is performed to ensure that the features are scaled appropriately. Finally, he extracted MFCC features are then converted into a NumPy array.

3.2 Proposed model

For the research work, we have used a custom model based on a Bi-Directional LSTM architecture used in conjunction with the MFCC feature extraction for audio feature learning. A Bi-Directional LSTM is a type of Recurrent Neural Network (RNN) which processes sequences in both forward and backward direction, which helps in learning more about the data by capturing both contexts of past and future. LSTM units are specialized RNN units capable of learning long-term dependencies. In this case, the Bi-directional LSTM captures complex patterns in the MFCCs.

The model consists of 7 layers. It has three Bi-LSTM layers with either 64 or 32 units. The LSTM layers are followed by strategically placed dropout layers with a dropout rate of 0.3 to prevent overfitting. Dropout regularization helps to improve the generalization of the model. Two Dense (fully connected) layers are included for feature learning and to perform classification based on the learned features. ReLU is commonly used in hidden layers to introduce non-linearity into the model. Softmax function converts the raw output scores into probabilities, ensuring that the sum of probabilities across all classes is equal to 1. This makes it suitable for multi-class classification problems. The model uses categorical_crossentropy as a loss function while using Adam (Adaptive Moment Estimation) as the optimization algorithm. Data is distributed as 75% training data, 15%

testing data, and the remaining 15% is used as validation data. The model is trained for 15 epochs each with a batch size of 32.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #   
=================================================================
 bidirectional (Bidirection  (None, 100, 128)          66048     
 al)                                                             
                                                                 
 dropout (Dropout)           (None, 100, 128)          0         
                                                                 
 bidirectional_1 (Bidirecti  (None, 100, 64)           41216     
 onal)                                                           
                                                                 
 dropout_1 (Dropout)         (None, 100, 64)           0         
                                                                 
 bidirectional_2 (Bidirecti  (None, 64)                24832     
 onal)                                                           
                                                                 
 dense (Dense)               (None, 32)                2080      
                                                                 
 dropout_2 (Dropout)         (None, 32)                0         
                                                                 
 dense_1 (Dense)             (None, 10)                330       
                                                                 
=================================================================
Total params: 134506 (525.41 KB)
Trainable params: 134506 (525.41 KB)
Non-trainable params: 0 (0.00 Byte)
```

Fig 3: Model Summary of Bi-Directional LSTM model

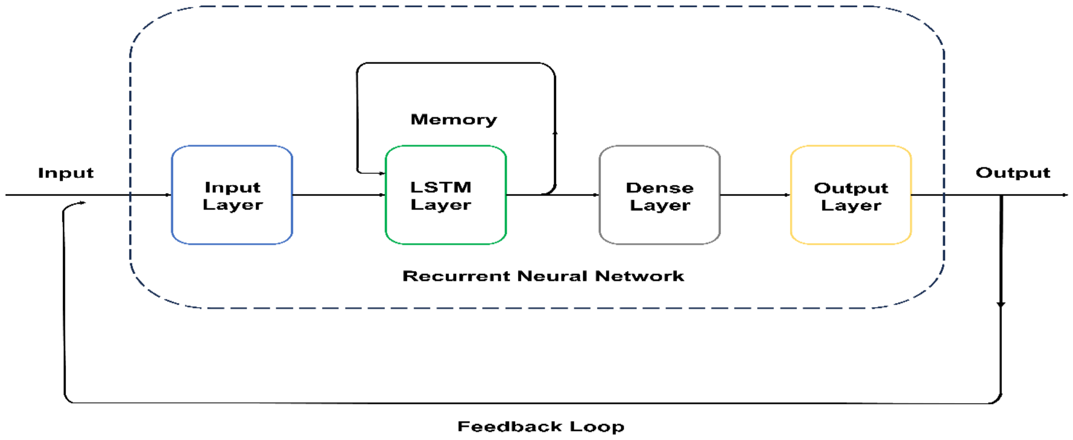


Fig 4: Feedback loop of LSTM model

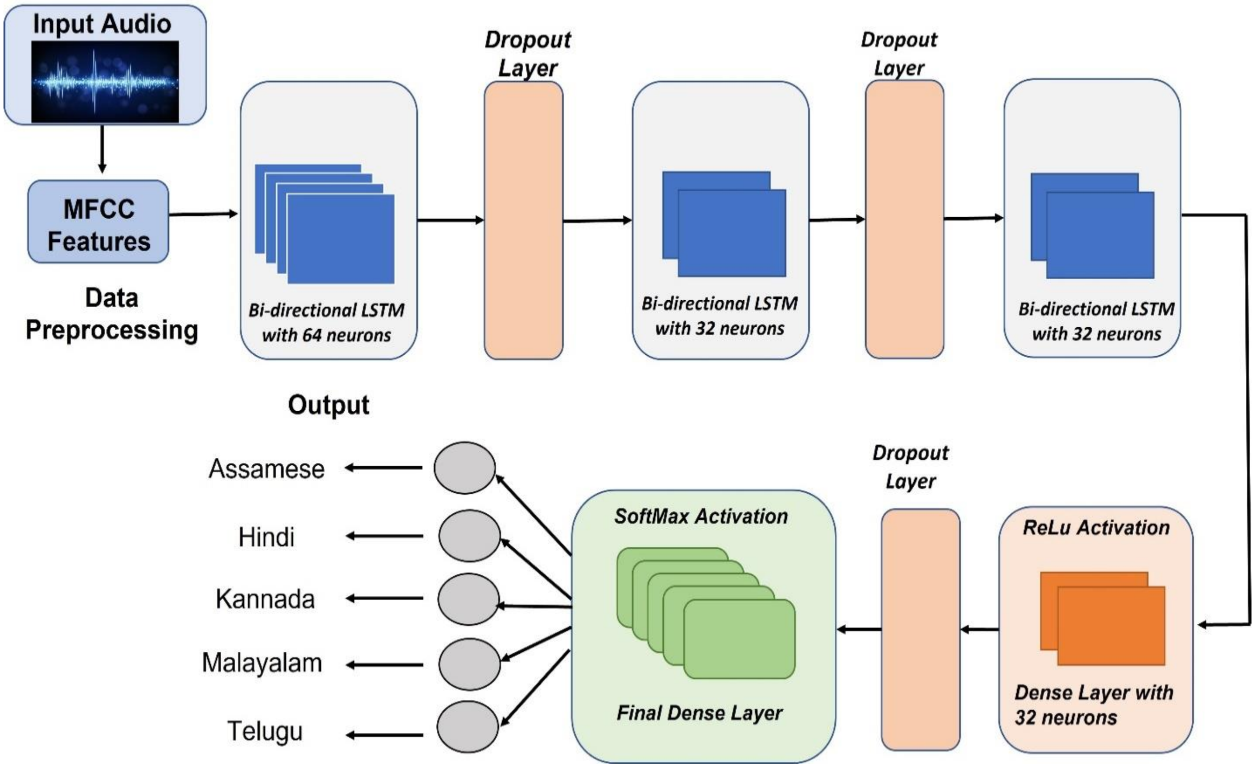


Fig 5: Model Architecture of Bi-Directional LSTM model

4. Experiments

In this work, the experiment is conducted in five languages namely, Assamese, Hindi, Kannada, Malayalam and Telugu. The model was trained for 15 epochs each with a batch size of 32 The Indian audio language detection model achieved impressive results. The model attained high accuracy across all evaluation stages, achieving a score between 95.6% and 99.1% accuracy. This indicates the model effectively learned to distinguish between different languages The following figures show the accuracy and loss of the model. The accuracy scores and loss scores are given below.

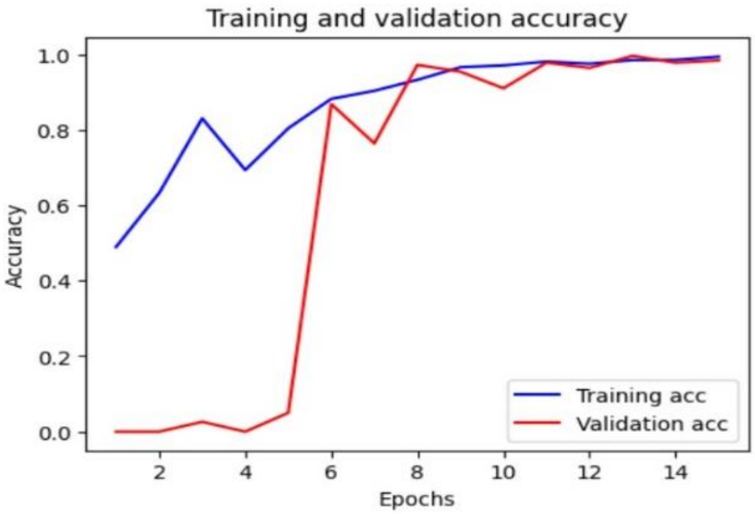


Fig 6: Loss graph of model

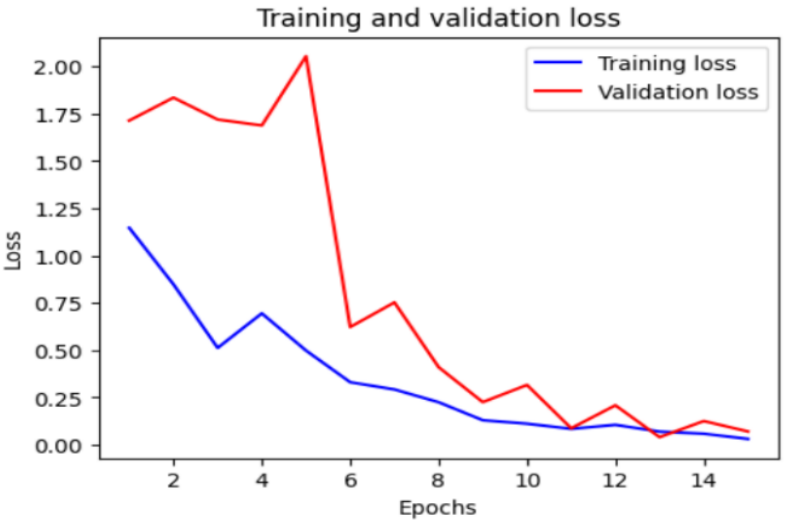


Fig 7: Accuracy graph of Model

5. Discussion and Conclusion

5.1 Discussion

In this work we have developed an Indian Language Identification system for five Indian language with the help of a deep learning approach. The LID system can be said to be the front end for various Automatic Speech recognition-based applications. The following figures shows the performance of the model on evaluation parameters using the confusion matrix.

Dataset	Accuracy	Loss
Training	98.50%	0.0576
Validation	95.60%	0.181
Testing	99.10%	0.205

Table 1: Model Performance on data split

We see the proposed LID system has performed well on various parameters and has been able to detect languages quite well. the model achieved a remarkable test accuracy ranging from 94% to 98%. This accomplishment highlights the effectiveness of two key elements within the model's architecture. Firstly, Mel-frequency cepstral coefficients (MFCCs) played a crucial role in extracting relevant features from the speech audio. By analyzing the spectral characteristics of the spoken language, MFCCs provided the model with the necessary information to distinguish between languages, as each possesses unique frequency patterns.

Secondly, the utilization of Bidirectional Long Short-Term Memory (LSTM) units proved highly beneficial. LSTMs' ability to learn long-term dependencies within speech sequences, combined with the bidirectional processing that analyzes audio in both directions, empowered the model to gain a deeper understanding of the language structure. Notably, the model achieved this impressive performance even when trained on low-resource hardware. This low-resource compatibility expands the model's potential reach, making it particularly valuable for deployment in areas with limited computational resources, such as rural India. The confusion matrix issued to evaluate the performance of the model across parameters such as Precision, Recall, F1 score, and Accuracy.

5.2 Comparison

The proposed model is compared to other models on Indian language audio identification/detection models that employ different architecture and feature extraction technologies. The results shows that the proposed model compares favorably with the other models, that were trained with state-of-the-art hardware and surplus training data, offering similar results and accuracy.

Model	Technology Used	Highest Accuracy achieved
Proposed Method (2024)	MFCC, Bi-Directional LSTM	99.10%
Ahmed et al. (2021)	Mel Spectrogram, CNN	100%
Basu et al (2021)	MFCC, SDC, RASTA-PLP +LSTM, GMM, VNN	95.69%
Godbole et al (2020)	Log Spectrogram, Gammatonegram, IIR-CQT Spectrogram, CNN	98.86%
Deshwal et al. (2020)	MFCC + RASTA-PLP, MFCC + SDC, Feed-Forward Neural Network	94.60%
Sisodia et al. (2020)	MFCC + Ensemble Methods	85.40%
Guha at el. (2020)	Mel Spectrograms, RASTA-PLP, NS, HS, NMR	99.89%

Table 2: Comparison of models and their accuracy figures.

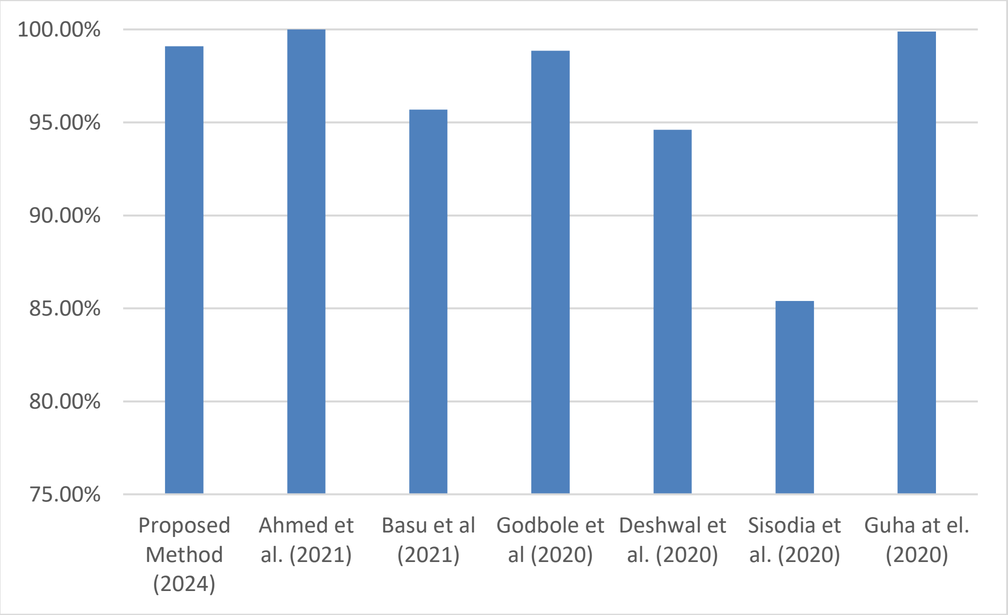
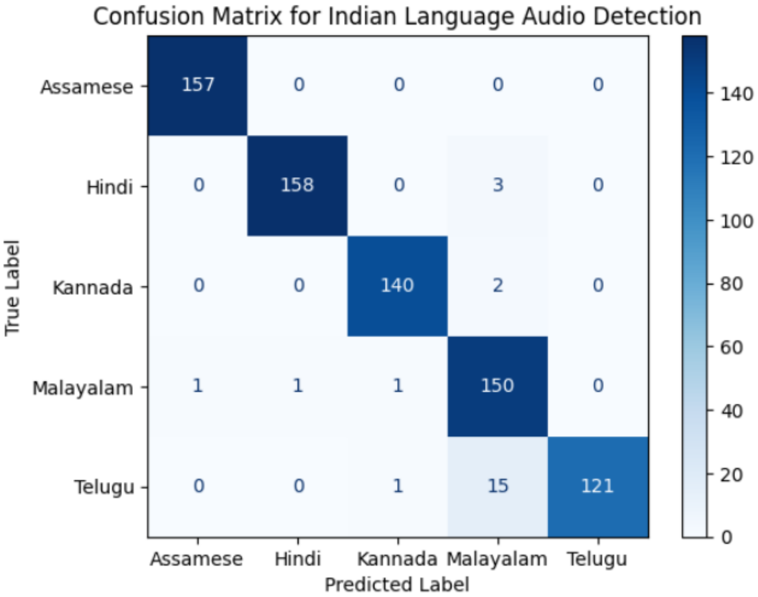


Fig 8: Bar chart comparison of the models

5.3 Evaluation Parameters

Given evaluation parameters are always desirable when assessing the performance of the given model.



True Label \ Predicted Label	Assamese	Hindi	Kannada	Malayalam	Telugu
Assamese	157	0	0	0	0
Hindi	0	158	0	3	0
Kannada	0	0	140	2	0
Malayalam	1	1	1	150	0
Telugu	0	0	1	15	121

Fig 9: Confusion Matrix of Data

Accuracy: It is the ratio of correctly predicted instances to total instances. It can be expressed as

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

Precision: It is the ratio of correctly predicted positive instances to total predicted positive outcomes. It can be expressed as

$$\frac{TP}{TP+TN}$$

Recall: It is the ratio of correctly predicted positive instances to the total predicted positives It can be expressed as

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

Where ***TP*** represents all true positives, ***TN*** denotes all true negatives, ***FP*** shows all false positives, ***FN*** represents all false negative

F1 Score: It is the mean of precision and recall.

It can be expressed as $\frac{2(P\times R)}{P+R}$

Metrics	Score
Accuracy	0.9873
Precision	0.968
Recall	0.968
F1 Score	0.968

Table 3: Performance parameters of model

6. Future Works

The research work in this paper describes a Language Identification for five Indian Languages namely Assamese, Hindi, Kannada, Malayalam and Telugu. With more available data, the model can be extended to encompass more high-demand Indian languages and dialects. The model's existing knowledge can be incorporated in newer models with the help of Transfer Learning techniques for faster training and implementation. Incorporating code switching capabilities i.e., recognizing multiple languages at a time to enhance model performance in multilingual speech scenarios. Integrating with an Automatic Speech Recognition (ASR) system would enable not only language identification but also speech-to-text conversion in the identified language in real time. This can lead to many real-world applications for the LID system. The LID system can be deployed in call centers, low end embedded systems, android and web apps.

References

[1] Deshwal, D., Sangwan, P., & Kumar, D. (2020). A Language Identification System using Hybrid Features and Back-Propagation Neural Network. *Applied Acoustics*, *164*, 107289. https://doi.org/10.1016/j.apacoust.2020.107289

[2] Saini S, Sahula V. Language learnability analysis of Hindi: a comparison with ideal and constrained learning approaches. J Psycholinguist Res 2019:1–14.

[3] Dey, S., Sahidullah, M., & Saha, G. (2022). An Overview of Indian Spoken Language Recognition from Machine Learning Perspective. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(6), 1–45. https://doi.org/10.1145/3523179

[4] S. Tong, P. N. Garner, and H. Bourlard, "An investigation of deep neural networks for multilingual speech recognition training and adaptation," in INTERSPEECH, pp. 714–718, ISCA, 2017.

[5] P. Jain, K. Gurugubelli, and A. K. Vuppala, "Towards emotion independent language identification system," in International Conference on Signal Processing and Communications (SPCOM), pp. 1–5, IEEE, 2020.

[6] P. Matějka, O. Novotny, O. Plchot, L. Burget, M. D. Sánchez, and J. Černock ` y, "Analysis of score normalization in multilingual speaker ` recognition," INTERSPEECH, pp. 1567–1571, 2017.

[7] Thukroo, I. A., & Bashir, R. (2021). Spoken Language Identification System for Kashmiri and Related Languages Using Mel-Spectrograms and Deep Learning Approach. *Spoken Language Identification System for Kashmiri and Related Languages Using MelSpectrograms and Deep Learning Approach*. https://doi.org/10.1109/icsc53193.2021.9673212

[8] Basu, J., Khan, S., Roy, R., Basu, T. K., & Majumder, S. (2021). Multilingual speech corpus in Low-Resource Eastern and Northeastern Indian languages for speaker and language identification. *Circuits, Systems, and Signal Processing*, *40*(10), 4986–5013. https://doi.org/10.1007/s00034-021-01704-x

[9] Godbole, S., Jadhav, V., & Birajdar, G. (2020). Indian Language Identification using Deep Learning. ITM Web of Conferences, 32, 01010. https://doi.org/10.1051/itmconf/20203201010

[10] D. S. Sisodia, S. Nikhil, G. S. Kiran and P. Sathvik, "Ensemble Learners for Identification of Spoken Languages using Mel Frequency Cepstral Coefficients," 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, 2020, pp. 1-5, doi: 10.1109/IDEA49133.2020.9170720.

[11] T. G. Dietterich, "Ensemble Methods in Machine Learning," Mult. Classif. Syst., vol. 1857, pp. 1–15, 2000

[12] Guha, S., Das, A., Singh, P. K., Ahmadian, A., Senu, N., & Sarkar, R. (2020). Hybrid feature selection method based on harmony search and naked Mole-Rat algorithms for spoken language identification from audio signals. IEEE Access, 8, 182868–182887. https://doi.org/10.1109/access.2020.3028121

[13] Rangan, Pradeep & Teki, Sundeep & Misra, Hemant. (2020). Exploiting Spectral Augmentation for Code-Switched Spoken Language Identification.

[14] Madhu, C., George, A., & Mary, L. (2017). Automatic language identification for seven Indian languages using higher level features. *Automatic Language Identification for Seven Indian Languages Using Higher Level Features*. https://doi.org/10.1109/spices.2017.8091332