Spoken Language Identification for Indo-Aryan And Dravidian Family

Anirban Ghosh[1*], Amal Kuniyil Parambath[1†], Arpita Bhattacharjee[1†] and Thoudam Doren Singh[1†]

[1*]Computer Science & Engineering Department, National Institute of Technology ,Silchar, Silchar, 788010, Assam, India.

*Corresponding author(s). E-mail(s): anirbangh669@gmail.com;
Contributing authors: amalkuniyilparambath@gmail.com;
arpitaofficial193@gmail.com; thoudam.doren@gmail.com;
†These authors contributed equally to this work.

Abstract

Spoken Language identification is the process of identifying the language being spoken in the audio. Investigation in the area of spoken language identification on regional languages aids in broadening the outreach of technology to regional language speakers and also gives to the preservation of regional languages. This paper studies the use of Convolutional Neural Networks for spoken language recognition of three local Indian languages Hindi, Bengali and Malayalam. Wave-Pad Sound Editor was used to create the data set. The samples are equally balanced between languages and genders of the speakers not to favor any subgroup. Attention has been paid to concentrating more on language properties than any specific voice. All the audio clips are the mono channel. The input audio is trimmed to 6s, 8s and 10s after which the MFCC spectrogram and filter banks are extracted. The pre-processed data is passed to the convolution neural network. The model generated in this method was tested against real-life clips and could generalize well. The high efficiency of the model can be attributed to the balanced distribution of the data set.

Keywords: spoken language identification, speech recognition, Convolutional neural networks, melSpectrogram

1 Introduction

Spoken language identification or automatic language identification is the process of identifying the language used by the speaker from the given digitized speech utterance. Language has played an important role in facilitating communication and in the exchange of ideas among people. When a multitude of languages is spoken in an environment, the first step in communication is the identification of the language used. The applications of language identification are several in number. In countries such as India, where numerous languages are spoken by the people, the existence of an automatic language identification system tends to serve as a means to simplify several existing processes. Consider the customer care center of a telecommunications company that has to answer calls from several customers to address their grievances or to provide information. In order to ensure efficient functioning, it is essential that the customer is able to express their problems through a comfortable means. Generally, when a customer calls a support center, they are first prompted to select the language they would like to use. Instead, if an automatic language identification system was employed, the customer could start explaining their problem or ask for any information right away without having to select an option to choose the language used for the call. In this manner, the average time required to complete each call can be reduced to a certain extent. Another important application of automatic language identification is speech-to-text conversion. When a computer is equipped with an automatic language identification system, it enables the user to simply speak instead of having to type commands. Automatic language identification systems are particularly useful when composing messages in different languages. Users need not have to manually change the language each time they wish to change the language of the content. The system will be able to identify the language used by the speaker and use the appropriate font to compose the message as it is dictated.

The main challenge of spoken language identification is finding meaningful audio feature representations which are robust to individual variations in pronunciation as well as to similarities of languages within the same language families. When the classifier is trained, it is able to discriminate different classes. The most often used acoustic features are extracted from a short segment of a speech signal, typically 60 to 80 nillis. The most commonly used features are Mel Frequency Cepstral Coefficient(MFCC), Linear Prediction Coding(LPC) , Linear Prediction Cepstral Coefficient(LPCC) and Perceptual Linear Prediction(PLP).

In this paper we first discuss about Dataset collection and preprocessing , MFCC and later use this as a feature extractor and train a CNN architecture to predict the language in a given audio clip. We have restricted the choice of languages to the Indo-Aryan(Hindi, Bengali) and Dravidian(Malayalam) family of languages.

2 Related Works

Language recognition is still far from perfect. As far as language characterization is concerned, we have not been able to effectively venture beyond acoustic– phonetic and phonotactic knowledge, despite the fact that there exists strong evidence in human listening experiments that prosodic information, syllable structure, and morphology are useful knowledge sources [9]. A lot of research has been done towards automatic language identification by combining different feature extraction methods and devising different approaches. Convolutional Neural Network (CNN) Model along with MFCC features ([1], [5]) gives good accuracy of above 90%. Using Convolutional Recurrent Neural Network (CRNN) , which is a hybrid of Recurrent Neural Network (RNN) and CNN, has been proven to increase the accuracy furthermore ([1], [8]). As of late, I-Vector based strategies have progressed toward becoming cutting edge in LID, firmly following comparative improvements in the Speaker Identification (SID). In [10], the fusion of bottleneck and posterior feature-based approaches with Deep Neural Networks (DNNs) trained with different languages results in systems that are between 40 and 70 % better than the baseline Gaussian Mixture Model (GMM) systems over all test durations.

All the more as of late, DNN and CNN based LID methodologies are ending up giving better execution analyzed than I-Vector based LID strategies figured utilizing a GMM Universal Background Model(GMM-UBM) based structure [2]. In [7], the system using MFCC features of speech input signal, along with Random Forest (RF), GMM, and K-Nearest Neighbor (KNN) as a classifier, using the 3s, 10s, and 30s as scoring method, with dataset that consists of Javanese, Sundanese, and Minang languages from Indonesia, has given an accuracy of 98.88%, 95.55% and 82.24% for KNN, RF and GM respectively for 30s of speech.

A lot of research has been done towards automatic language identification of Indian languages by combining different feature extraction methods and devising different approaches. In [3], SLID for Indian languages in spoofing and noisy environments using a concatenated CQCC-MFCC (Constant Q Cepstral Coefficients) feature set applied to a CNN It was found to have a 97% accuracy on the INDIC TTS Database. In [4], SLID for four local Indian languages Kannada, Hindi, Tamil and Telugu using MFCC features along with Support Vector Machines (SVM) and Decision Tree (DT) classifiers gave an accuracy of 76% and 73% respectively. In [6], using filter banks instead of MFCC with CNN has given an accuracy of 92.74%.

3 Dataset and Features

3.1 Dataset

The dataset contains Hindi, Bengali and Malayalam audio recordings. The audio samples were taken from Mann ki baat and All India Radio. The advantage of using this dataset is there are several speakers of different gender

available which helps to generate a generalized model. Each clips are changed into .flac documents rapidly to stay away from re-encoding(and losing quality) during changes. Each example is a .flac sound record with a sample rate of 44.1kHz , bit depth of 16, single channel. The audio clips were trimmed into 6s, 8s and 10s.

Bengali clips were collected from 7 videos of Mann Ki Baat of total duration of 3 hours from All India Radio Kolkata. Hindi clips were collected from 14 videos of daily Hindi news of total duration 3 hours from All India Radio. Malayalam clips were collected from 19 videos of daily Malayalam news of total duration 3 hours from All India Radio Thiruvananthapuram.

Dataset	10s	8s	6s
Training	3000	3750	5000
Validation	332	450	600
Testing	173	250	300

Table 1: Statistics of the entire data

Language	10s	8s	6s
Hindi	1141	1444	1915
Bengali	1223	1546	2049
Malayalam	1141	1460	1936

Table 2: Statistics of data for each language

Hindi	10s	8s	6s
Training	1000	1187	1614
Validation	84	150	200
Testing	58	84	100

Table 3: Data statistics for Hindi

Bengali	10s	8s	6s
Training	1000	1359	1749
Validation	164	150	200
Testing	58	83	100

Table 4: Data statistics for Bengali

Malayalam	10s	8s	6s
Training	1000	1359	1749
Validation	84	150	200
Testing	57	83	100

Table 5: Data statistics for Malayalam

3.2 Features

MFCC: Mel frequency cepstral coefficients (MFCC) have proved to be one of the most successful feature representations in speech related recognition tasks.Audio signals are composed of a set of contrasting frequency components each of which is found in different proportions in it. This set of components along with their frequencies is called the spectral envelope. It is this envelope that plays a significant role in determining what human beings actually hear, and this envelope is very well represented by spectrogram. It is a representation of frequencies changing with respect to time for given music signals. However, to model human hearing more accurately, mel-scale spectrogram have been used so that two pairs of frequencies separated by a delta in the mel scale are perceived by humans as being equidistant. The mel scale is nothing but a non-linear transformation of frequency scale based on the perception of speeches. The mel frequency scale for frequency f in Hertz is defined by equation 1 and its inverse is defined by equation 2.

$$mel = 2595 \log_{10}(1 + \frac{f}{700}) \tag{1}$$

$$f = 700(10^{\frac{mel}{2595}} - 1) \tag{2}$$

To compute mel-spectrogram of a signal, the speech signal is first sampled into separate windows by applying framing and windowing techniques. The Fast Fourier Transform (FFT) of each frame is computed to transform from time domain to frequency domain. Now a mel scale is generated by taking the entire frequency spectrum and separating it into evenly spaced frequencies. Finally, the mel-spectrogram is generated for each window by decomposing the magnitude of the signal into its components, corresponding to the frequencies in the mel scale. Figure 1 summarizes all the processes and steps taken to obtain the mel-spectrogram coefficients.

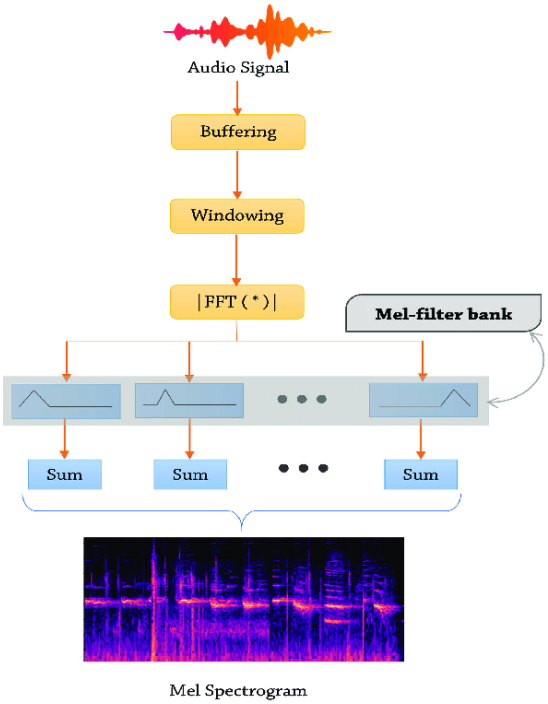

Fig. 1: Steps involved in mel-spectrogram feature extraction.

LPC: Linear Predicative Coding (LPC) is a popular approach for modeling human voice production which performs well in clean environment but not so in noisy. The parameters for speech signals are: Pitch Period, Speech Frame Energy, Formant, Short Time Spectrum and bandwidth. LPC is a very important feature used in auditory and processing of speech signal which draws out parameters of speech like spectra and pitch formants. LPC is also known as temporal approach which is invented to make it equivalent to the resonant structure of human vocal tract that developed the subsequent sound. The methodology behind the usage of LPC is to reduce the squared difference between original and estimated speech signal at a finite time. The block diagram of LPC is depicted in Figure 2. Steps for the computation of LPC are as follows:

1. Pre-Emphasis: Firstly the analysis of speech sample is done by passing it through a filter with the goal of spectrally flatten the signal and to make it less prone to precision effect. The coefficient of filter should be between 0.9 and 1.
2. Framing: After pre-emphasis step, the resulting speech are divided into frames consists of M samples each of 20 to 40 second. There is a standard overlap of 10ms between two adjacent frames to ensure stationary between frames.

3. Windowing: In this, the resulting frames are multiplied by hamming window with the purpose of minimizing the edge effect.
4. Computing the LPC: In the final step, the method of auto-correlation is functional upon the frames of windowing speech sample. Maximum Autocorrelation value is when analysis of the order of LPC is evaluated.

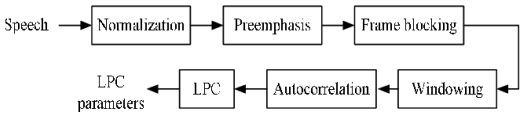

Fig. 2: Block diagram of LPC

PLP: Perceptual linear prediction (PLP) technique combines the critical bands, intensity-to-loudness compression and equal loudness pre-emphasis in the extraction of relevant information from speech. It is rooted in the nonlinear bark scale and was initially intended for use in speech recognition tasks by eliminating the speaker dependent features. It is similar to LPC except from the fact that PLP works in close resemblance to that of the human auditory system. PLP gives a representation conforming to a smoothed short-term spectrum that has been equalized and compressed similar to the human hearing making it similar to the MFCC.

In the PLP approach, several prominent features of hearing are replicated and the consequent auditory like spectrum of speech is approximated by an autoregressive all–pole model. PLP gives minimized resolution at high frequencies that signifies auditory filter bank based approach, yet gives the orthogonal outputs that are similar to the cepstral analysis. It uses linear predictions for spectral smoothing, hence, the name is perceptual linear prediction. PLP is a combination of both spectral analysis and linear prediction analysis. The block diagram of PLP is depicted in Figure 3. The power spectrum of windowed signal is calculated as,

$$P(\omega) = Re(S(\omega))^2 + Im(S(\omega))^2 \tag{3}$$

A frequency warping into the Bark scale is applied. The first step is a conversion from frequency to bark, which is a better representation of the human hearing resolution in frequency. The bark frequency corresponding to an audio frequency is,

$$\Omega(\omega) = 6ln\left[\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right] \tag{4}$$

The auditory warped spectrum is convoluted with the power spectrum of the simulated critical-band masking curve to simulate the critical-band integration of human hearing. The smoothed spectrum is down-sampled at intervals of

≈ 1 Bark. The three steps frequency warping, smoothing and sampling are integrated into a single filter-bank called Bark filter bank.

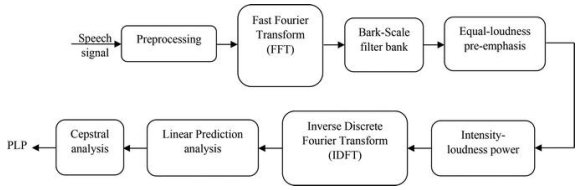

Fig. 3: Block diagram of PLP

4 Neural Network Architecture

The solution is based on Convolutional Neural Network(CNN) in order to detect language specific phonemes. It supports 3 languages: Hindi, Bengali and Malayalam. The preprocessed data is passed to the Convolutional Neural Network. Table 1 summarizes the layer-wise model architectures which has seven blocks each consisting of 2D convolutional layers, batch normalization, and max pooling for downsampling along time and frequency. The learning of more abstract features in higher layers is favoured by increasing the number of kernels from 32 to 512 from the first to the last convolutional layer.

First layer or the input layer takes as input, a numpy array of shape 40x1000x1(since, all the filter banks generated were grouped together based on the speakers of the audio samples and numpy arrays were generated to represent these filter banks efficiently). This model includes five convolutional layers. For each of the convolutional layers, 2D Max Pooling of strides (2,2) is applied, decreasing the overall input size for next hidden layers by a factor of 2. Now, to minimize over fitting by reducing the total number of parameters in the model, Max Pooling is applied. The output is then flattened and fed into a simple layer consisting of 256 nodes. All the above layers use ReLU as their activation functions other than the output layer which consists of 3 nodes and uses softmax as its activation function. Fig. (insert no.) gives a schematic representation of the model showing interconnection between each layer and Table 6 gives an overview of each layer in the proposed model.

Layer	Channels	Kernel Size	Padding
ConvBlock2D-1	32	7	3
ConvBlock2D-2	64	5	2
ConvBlock2D-3	128	3	1
ConvBlock2D-4	256	3	1
ConvBlock2D-5	512	3	1
Dense	256		
Dense	N		

Table 6: Layer-wise model architecture for the CNN model

5 Training Process

Due to the large size of the dataset, the dataset was zipped. Now we fed this dataset to the model for training. The number of epochs was set to 100. During each epoch, 32 audio files were selected in random. The training data for each epoch is stored in a tensor with 40 mel spectrogram patches. The Adam optimizer was used for learning and loss was calculated using SoftmaxCrossEntropyLoss.

6 Experiments

Initially our model was trained on a dataset of three Indian languages. Each audio clip is of three seconds duration in training and testing.
Categorical Cross Entropy Loss is calculated as

$$loss = -\sum_{i=1}^{n} t_i log(p_i) \tag{5}$$

where n is the number of classes,t_i is the truth label and p_i is the Softmax Probability for the i^{th} class.
The following formula is adopted to calculate accuracy.

$$accuracy = \frac{\text{Number of correct predictions}}{\text{Total Number of predictions}} \tag{6}$$

7 Results

The goal was to recognise language among 3 target languages: Hindi(HI), Bengali(BA) and Malayalam(MA). Cross validation is used by splitting the whole dataset into 10 folds and the number of instances of each class is balanced in each fold, based on the speakers of the speech samples, their gender and the languages they speak. This provides a robust estimate of the performance of the model on unseen data. Accuracy was found out by testing out all models with 100 clips of each language of random duration and gender.

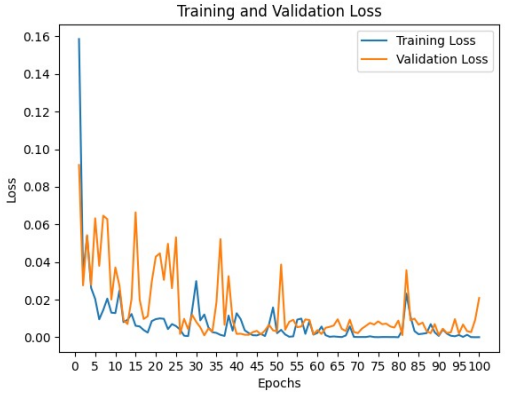

Fig. 4: Variation of Model Loss with epochs for 6 seconds

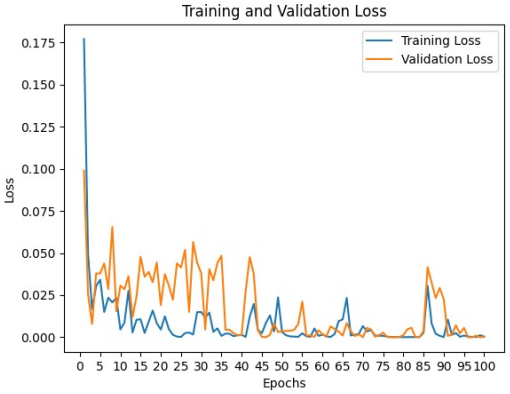

Fig. 5: Variation of Model Loss with epochs for 8 seconds

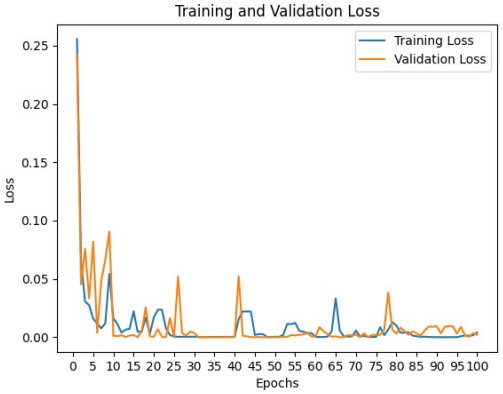

Fig. 6: Variation of Model Loss with epochs for 10 seconds

Model	Accuracy		
	Features		
	MFCC	LPC	PLP
6s_model	97.93%	91.56%	97.23%
8s_model	98.62%	94.33%	98.06%
10s_model	99.31%	94.89%	97.51%

Table 7: Accuracy Table

8 Conclusion

Our results show that convolutional neural networks combined with mel spectrogram representations of speech signals are an adequate processing pipeline for the given task. Languages are not easily treated as discrete, identifiable units with precise boundaries between them. This is especially valid for languages coming from the same language family such as Indo- Aryan or Dravidian. Moreover, every language is characterized by variations between the communities that use it, which often leads to a variety of local accents and dialects. Given a potential application scenario of a larger number of languages to be identified, a hierarchical classification approach could be beneficial. Here, the language family should be classified first. Then, specific classification models, which are optimized towards languages of certain families, will likely perform better than a general-purpose language classifier.

This paper surveys a comparative analysis of 6s, 8s and 10s SLID models using MFCC, LPC and PLP feature extraction technique and classifier as CNN. As seen from table 7, the accuracy of the 10s model was found out to

be the highest, followed by 8s model. Also using MFCC as feature extraction technique yielded the highest accuracy.

The Model made available can be scaled to support many more languages with the availability of relevant dataset. One key aspect that can be explored is the use of this model to classify languages having accent dependencies. Scope for improvement lies in the fact, that this is a simple CNN implementation. Hence, this model can be extended to make use of LSTMs and other RNN related deep-learning models to get better results. Finally, combining all these technologies can result in support for languages with low resources which can be used in tasks such as accent and dialect detection which is still a key area for research exploration.

References

[1] Draghici, Alexandra & Abeßer, Jakob & Lukashevich, Hanna. (2020). A Study on Spoken Language Identification using Deep Neural Networks. 10.1145/3411109.3411123.

[2] B. Aarti and S. K. Kopparapu, "Spoken Indian language classification using artificial neural network — An experimental study," 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2017, pp. 424-430, doi: 10.1109/SPIN.2017.8049987.

[3] A. R. Ambili and R. C. Roy, "Spoken Language Identification of Indian Languages in Adversarial Synthetic and Noisy Attacking Environments," 2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS), Kochi, India, 2022, pp. 1-6, doi: 10.1109/IC3SIS54991.2022.9885560.

[4] H. Venkatesan, T. V. Venkatasubramanian and J. Sangeetha, "Automatic Language Identification using Machine learning Techniques," 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2018, pp. 583-588, doi: 10.1109/CESYS.2018.8724070.

[5] L. R. Arla, S. Bonthu and A. Dayal, "Multiclass Spoken Language Identification for Indian Languages using Deep Learning," 2020 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2020, pp. 42-45, doi: 10.1109/IBSSC51096.2020.9332161.

[6] S. Mukherjee, N. Shivam, A. Gangwal, L. Khaitan and A. J. Das, "Spoken Language Recognition Using CNN," 2019 International Conference on Information Technology (ICIT), Bhubaneswar, India, 2019, pp. 37-41, doi: 10.1109/ICIT48102.2019.00013.

[7] Wicaksana, Vincentius S.Kom, Amalia. (2021). Spoken Language Identification on Local Language using MFCC, Random Forest, KNN, and GMM.

International Journal of Advanced Computer Science and Applications. 12. 10.14569/IJACSA.2021.0120548.

[8] A. Alashban, M. Qamhan, A. Meftah, and Y. Alotaibi, Spoken language identification system using convolutional recurrent neural network, Sep. 2022. DOI: 10.3390/app12189181.

[9] H. Li, B. Ma and K. A. Lee, "Spoken Language Recognition: From Fundamentals to Practice," in Proceedings of the IEEE, vol. 101, no. 5, pp. 1136-1159, May 2013, doi: 10.1109/JPROC.2012.2237151.

[10] L. Ferrer, Y. Lei, M. McLaren and N. Scheffer, "Study of Senone-Based Deep Neural Network Approaches for Spoken Language Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 1, pp. 105-116, Jan. 2016, doi: 10.1109/TASLP.2015.2496226.

[11] Bakshi, Aarti & Kopparapu, Sunil Kumar. (2021). Improving Indian Spoken-Language Identification by Feature Selection in Duration Mismatch Framework. SN Computer Science. 2. 10.1007/s42979-021-00750-1.

[12] Das, Himanish Roy, Pinki. (2020). Bottleneck Feature-Based Hybrid Deep Autoencoder Approach for Indian Language Identification. ARABIAN JOURNAL FOR SCIENCE AND ENGINEERING. 45. 3425–3436. 10.1007/s13369-020-04430-9.

[13] Bhanja, Chuya Laskar, Azharuddin & Laskar, Rabul. (2019). A Preclassification-Based Language Identification for Northeast Indian Languages Using Prosody and Spectral Features. Circuits Systems and Signal Processing. 38. 10.1007/s00034-018-0962-x.

[14] S. Guha, A. Das, P. K. Singh, A. Ahmadian, N. Senu and R. Sarkar, "Hybrid Feature Selection Method Based on Harmony Search and Naked Mole-Rat Algorithms for Spoken Language Identification From Audio Signals," in IEEE Access, vol. 8, pp. 182868-182887, 2020, doi: 10.1109/ACCESS.2020.3028121.

[15] M. -G. Wang, Y. Song, B. Jiang, L. -R. Dai and I. McLoughlin, "Exemplar based language recognition method for short-duration speech segments," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 7354-7358, doi: 10.1109/ICASSP.2013.6639091.

[16] A. Poddar, M. Sahidullah and G. Saha, "Performance comparison of speaker recognition systems in presence of duration variability," 2015 Annual IEEE India Conference (INDICON), New Delhi, India, 2015, pp. 1-6, doi: 10.1109/INDICON.2015.7443464.

[17] Jothilakshmi, S. & Ramalingam, Vivekanandan & Palanivel, Sengottayan. (2012). A hierarchical language identification system for Indian languages. Digital Signal Processing. 22. 544–553. 10.1016/j.dsp.2011.11.008.

[18] Arruti, Andoni & Cearreta, Idoia & Álvarez, Aitor & Lazkano, Elena & Sierra, Basilio. (2014). Feature Selection for Speech Emotion Recognition in Spanish and Basque: On the Use of Machine Learning to Improve Human-Computer Interaction. PloS one. 9. e108975. 10.1371/journal.pone.0108975.

[19] Chowdhury, Amit & Borkar, Vaibhav & Birajdar, Gajanan. (2019). Indian language identification using time-frequency image textural descriptors and GWO-based feature selection. Journal of Experimental & Theoretical Artificial Intelligence. 32. 1-22. 10.1080/0952813X.2019.1631392.

[20] D. Sengupta and G. Saha, "Automatic recognition of major language families in India," 2012 4th International Conference on Intelligent Human Computer Interaction (IHCI), Kharagpur, India, 2012, pp. 1-4, doi: 10.1109/IHCI.2012.6481844.

[21] Koolagudi, Shashidhar & Bharadwaj, Akash & Srinivasa Murthy, Y.V. & Reddy, Nishaanth & Rao, Priya. (2017). Dravidian language classification from speech signal using spectral and prosodic features. International Journal of Speech Technology. 20. 10.1007/s10772-017-9466-5.

[22] S. Ranjan, C. Yu, C. Zhang, F. Kelly and J. H. L. Hansen, "Language recognition using deep neural networks with very limited training data," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 5830-5834, doi: 10.1109/ICASSP.2016.7472795.

[23] Hashem, Ahmad & Fakhr, Mohamed & Abdou, Sherif. (2011). Spoken Language Identification Using Ergodic Hidden Markov Models.

[24] Zazo, Ruben & Lozano-Diez, Alicia & Gonzalez-Dominguez, Javier & Toledano, Doroteo & Gonzalez-Rodriguez, Joaquin. (2016). Language Identification in Short Utterances Using Long Short-Term Memory (LSTM) Recurrent Neural Networks. PloS one. 11. e0146917. 10.1371/journal.pone.0146917.

[25] Nandi, Dipanjan & Rao, K.. (2015). Language Identification Using Excitation Source Features. 10.1007/978-3-319-17725-0.