Assamese Dialect Identification System using Convolution Neural Networks

Hem Chandra Das[1], Kshirod Sarmah[2*], Deepak Hajoary[3], Raju Narzary[4], Rinku Basumatary[5]

[1]Department of Computer Science and Technology, Bodoland University, Kokrajhar, 783370, Assam, India, hemchandradas78@gmail.com
[2]Department of Computer Science, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya (A Govt. Model College), Goalpara, 783124, Assam, India,
kshirodsarmah@gmail.com
[3]Department ofManagement Studies, Bodoland University, Kokrajhar, 783370, Assam, India, hajoary.deepak@gmail.com
[4]Department of Computer Science and Technology, Bodoland University, Kokrajhar, 783370, Assam, India, rajnarz@gmail.com
[5]Department of Computer Science and Technology, Bodoland University, Kokrajhar, 783370, Assam, India, rinkubasumatary@gmail.com

Corresponding Author: Kshirod Sarmah
*Department of Computer Science, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya (A Govt. Model College), Goalpara, 783124, Assam, India, kshirodsarmah@gmail.com

Abstract: Labeling speech in an audio file with appropriate dialect labels is the aim of a dialect identification system. This paper presents a method of using convolution neural networks (CNN) to identify four Assamese dialects: Goalporia dialect, Kamrupi dialect, Eastern Assamese dialect, and Central Assamese dialect. This study employed the speech patterns of four major Assamese regional dialects: the Central Dialects spoken in and around the district of Nagaon; the Eastern Assamese dialect spoken in the districts of Sibsagar and its neighboring areas; the Kamrupi dialect spoken in the districts of Kamrup, Nalbari, Barpeta, Kokarajhar, and some areas of Bongaigaon; and the Goaplari dialect spoken in the Goaplara, Dhuburi, and a portion of Bongaigaon district. Over the course of two hours, audio samples from each of the four dialects were used to train the classifier. Mel spectrogram pictures, which are produced from two to four second divisions of raw audio input with varying audio quality, are used by the CNN. The system's performance is also analyzed in relation to the lengths of the train and test audio samples. The proposed CNN model achieves an accuracy of 90.82 percent, which may be the best when compared to machine learning models.

Keywords: ADID, Mel Spectrogram, Classification, Assamese Dialect, Machine Learning.

1.0 INTRODUCTION

Dialect identification is currently a hot topic in signal processing research since it has so many uses in day-to-day life. Dialect identification is the process of identifying languages from spoken utterances. The accurate and timely identification of dialect from audio at different sample rates and noise levels is one of the challenges in this endeavor. Differentiating between dialects that sound identical over short stretches of time is another challenge [1].

The majority of people who are native to Assam, a state in northeastern India, as well as some areas of its surrounding states, including Meghalaya, Nagaland, and Arunachal Pradesh, speak the Assamese language, which is descended from the Indo-Aryan language family. Schedule-VIII of the Indian Constitution lists Assamese as a Major Indian Language. It is the official language of the state of Assam. The Bodo and Kacharis, two of Assam's early populations, had a significant impact on the lexicon, phonology, and grammar of the Assamese language, which originated from Sanskrit [2].The Assamese language is divided into four main dialect groupings, per contemporary study. The Central group is spoken in and around the modern-day Nagaon district and its neighboring territories, whereas the Eastern group is spoken in the Sibsagar district and its surroundings. Spoken in unincorporated Kamrup, Nalbari, Barpeta, Darrang, and part of Bongaigaon, is the Kamrupi dialect. The Goalparia group resides in Goalpara, Dhubri, and parts of the districts of Kokrajhar and Bongaigaon [2]. But as of right now, most people agree that Central Assamese is the primary or standard dialect [3]. Regarding the translation of Assamese dialects, no noteworthy research has been published.According to linguistics, a dialect is a socially different language used by a particular group of native speakers who share a similar vocabulary, syntax, and pronunciation pattern [4]. One aspect of human intelligence is the capacity to distinguish between spoken languages [5] [6]. Dialect identification is the initial step in creating a voice recognition system for any language that is dialect-independent. The CNN models have produced encouraging results in the last few decades [7].

In order to accomplish this goal, a number of deep neural network models were examined because it has already been noted that these architectures produce positive outcomes [8]. [9]. These model families can

automatically extract important features, but they all need to undergo an audio to frequency domain preprocessing step.
In this work, An Assamese dialect Identification (ADID) System has been developed to identify Assamese dialects by using deep learning techniques like convolutions and maxpoolings. In order to facilitate training and testing quickly and easily, a basic model (a Convolution neural network with fewer parameters) was used for this purpose. For training or validation, the suggested model transforms a batch of unprocessed audio data into a batch of melspectograms.

2.0 LITERATURE REVIEW

The science of dialect identification was founded in 1877 by George Wenker, who conducted a number of investigations to identify dialect regions [10]. Baily was among the first to recognize and designate the Midland dialect as a separate dialect. The study's conclusions led to the conclusion that dialects shouldn't be divided solely into classes or groups based on vocabulary because vocabulary might differ greatly within a given geographic location [11]. Another attempt to ascertain if the Midland region qualifies as a distinct dialect region was made by Davis and Houck [12]. Eleven cities along the north-south route had their phonological and lexical characteristics successfully extracted by the researchers [13].
The Arabic dialect was examined by Diab et al. [14] and Watson [15], who also enumerated its traits, established a connection between the Standard language and its regional variants, and categorized the main regional dialects. GMM is used by Ibrahim et al. [16] to distinguish between Arabic dialects based on prosodic and spectral traits. Combining spectral and prosodic features was found to improve recognition accuracy of the Malaysian Quranic dialect by 5.5 to 7%. There was a reported range of accuracy for prosodic parameters such pitch and duration as well as MFCC, ranging from 81.7% to 89.6%.
Numerous Indian languages have been the subject of dialect recognition studies. Regional Telugu dialects were distinguished using GMM and HMM by Shivaprasad and Sadanandam [17]. For this reason, the writers created a database of Telugu dialects. MFCC and its derivatives, including ∆MFCC and ∆∆MFCC characteristics, were used for recognition. From every spoken phrase, the study extracts 39 feature vectors, which it then uses the GMM and HMM models to analyze. The HMM model is underperformed by the GMM model. Still, other words with the same auditory properties were not identified.
Using only prosodic traits and a few lines from each dialect, Chittaragi and Koolagudi [18] identify Telugu dialects using the closest neighbor method. Authors who only took prosodic elements into account achieved 75% accuracy.Chittaragi et al. [19] used prosodic and spectral properties to identify five Kannada dialects. To identify dialects, the authors employed neural networks and SVM. In the shortest amount of time, the Neural Network produces good results using text-independent data. Five Hindi dialects were distinguished by K. S. Rao et al. [20] using spectral and prosodic features: Chattisgharhi, Bengali, Marathi, General, and Telugu. The ten (10) people in their database, each speaking for five to ten minutes on their own, speak for a total of one to one and a half hours.
Bakshi et al. created an Artificial Neural Network (ANN)-based language distinguishing classifier. Bakshi et al. created an Artificial Neural Network (ANN)-based language distinguishing classifier [21]. One hundred speech samples with duration of 5 minutes and a sampling rate of 16 KHz were included in the database. Tamil, Malayalam, Assamese, Gujarati, Hindi, Bengali, Marathi, Kannada, and Telugu were among the nine Indian languages employed in their research. The testing accuracy obtained with 13-MFCC, 13 ∆MFCC, and 13-∆∆MFCC feature descriptors was 37.9998 percent with window duration of 20msec and 42.666 percent with window duration of 100msec. According to their findings, increasing the window size did not result in a significant increase in accuracy. Because the languages were not separated based on family, but rather based on multi-language discrimination.
Madhu et al. [22] employed an ANN classifier to detect seven Indian languages, including Bengali, Telugu, Urdu, Hindi, Assamese, Manipuri, and Punjabi using a 2-hour speech database,. Using prosodic and phonotactic characteristics, they achieved 72% and 68% accuracy, respectively.
Deep Neural Network with Attention (DNN-WA) and the i-vector system were the two approaches used by Veera et al. [23] to compare Language Identification systems. The thirteen Indian languages that were included in the sample included Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Manipuri, Marathi, Odiya, Punjabi, Tamil, Telugu, and Urdu. The sample consisted of thirty hours of test data and fifteen hours of training data. To develop a LID system, both the MFCC and the Residual Cepstral Coefficients (RCC) were subjected to Shifted Delta Cepstral coefficients (SDC). EERs of 9.93% and 6.25%, respectively, were obtained using MFCC and RCC features. Fusing the attributes from both features also resulted in an EER of 5.76 percent using a technique called late fusion. According to their findings, switching from i-vector to DNN resulted in little or no improvement; however DNN-WA resulted in a significant improvement.
Moreno et al. proposed that a deep neural network can identify the language of a spoken utterance based on brief temporal acoustic characteristics [24]. They conducted their investigation using two datasets: the Google 5M LID corpus and the NIST Language Recognition Evaluation (LRE) 2009. An i-vector model and this DNN-based acoustic model's performance were contrasted. When working with vast volumes of data, the DNN model performed better than the i-vector-based method.

Boussard et al. [25] investigated LID utilizing phone calls dataset of eleven languages with a duration spanning from a few seconds to over a minute, using the OGI Multilanguage Corpus. Numerous classification techniques were used, such as frame-level data aggregation, Feed-Forward Neural Networks, Recurrent Neural Networks (RNN), CNN, and Gaussian Mixture Models (GMM), to generate predictions either directly on the call level or from the derived features such as MFCC, SDC, and spectral centroids. The best results were obtained when GMM and SDC features were combined, according to the results.

3.0 EXPERIMENTAL SETUP

3.1 Speech Database

Review of the current literature reveals that there is no standard database for the Assamese language and its dialects. A new database has been created with speech samples from all the dialect groups. The speech data consist of speech samples from 10 speakers (5 male and 5 female) representing each dialect regions have been recorded. A phonetically rich script was prepared to record the speech samples. The same script was used to record all the dialects, including the standard Assamese. The recording has been done at 16 KHz sampling frequency, 16-bit mono resolution. Subjective listening test of the recordings has been done using listeners from the respective dialect groups who were not involved in the recording process. The dataset comprise more than 10000 spoken utterances of both male and female native speakers. The statistical representation of speech database is shown in Table 1.

3.2 Feature Extraction

In speech recognition tasks, Mel frequency cepstral coefficients (MFCC) have proven to be one of the most successful feature representations. The mel-cepstrum makes use of auditory concepts as well as the cepstrum's decorrelating characteristic [26]. The audio data is being converted from wav to melspectograms. Because we were making spectrograms from audio data, we translated it to the mel scale, which resulted in "melspectrograms". For the purposes of this paper, these images will be referred to as "spectrograms." We use the formula to convert from f hertz to m mels.The equation for mel spectrogram is given below:

$$m = 2595log_{10}(1 + f/700) \qquad (1)$$

Melspectogram are comparable to images in that it is the graphical representations of sound data. CNNs outperform Machine Learning Models and Artificial Neural Networks in terms of performance analysis [27]. As a result, we trained our classifier using Deep Learning techniques such as Convolutions and Maxpoolings.
For the training process, we employed audio tracks that lasted two to four seconds. Our algorithm only needs a two to four second audio sample to forecast the dialect because it was trained on two-second audio files. Consequently, the dialect detection process is incredibly fast. In Figure 1, the intricate architecture is seen. Mel spectrograms are created from the raw audio signals, which enable overfitting to be avoided. These spectrograms are loaded into our proposed model, which is based on AlexNet, to obtain dialect identification.

Table 1. Statistical representation of the speech database.

Number of Speakers	10 (Five male and Five female) for each dialect group
Number of sessions	02
Intersession interval	At least one week
Data Types	Speech
Types of Speech	Read speech
Sampling Rate	16 KHz
Sampling format	Mono-channel,16bit resolution
Speech Duration	Each speaker is recording is for minimum 30 minutes in each session.
Microphone	Zoom H4N Portable Voice Recorder microphone
Acoustic Environment	Laboratory
Total duration of speech data	Minimum 10 hours for each dialect, including standard Assamese

Four dialects with standard Assamese datasets were segmented into audio clip of each clip lasting between 2-4 seconds. WAV is the file type that we use for all of our audio files. To organize our data in a logical manner, we utilized torch.utils.data.dataset. We set the time of each audio file to 2 seconds in our final dataset so that we can generate a large amount of data from the available data. At 16 KHz, each audio signal was sampled. We collected audio clips of Assamese Language including Eastern group, Central Group, Kamrupi, and Goalporia. Each dialects speaker is of various genders and has different accents. Each audio file is transformed to melspectograms during training. Melspectograms were created with Pytorch'storchaudio library [28]. Figure 2 shows a melspectogram of a Goalporia dialect audio file with duration of 3 seconds, without any transformations or augmentations. Time masking and frequency masking are the transformations used here.

Frequency masking refers to the application of masking to a spectrogram in the frequency domain, while time masking refers to the application of masking to a spectrogram in the time domain [29].
Table 2 summarizes the statistics of recorded speech database collected from each four dialects. The available data is presented in terms of hours, with a total of 5-6 hours of data from each of the four dialects.

Table 2. Data duration in hours.

Dialects	Training Data Duration	Testing Data Duration
Eastern Dialects	4.30	0.83
Central Dialects	6.21	1.74
Kamrupia Dialects	5.76	1.35
Goalporia Dialects	4.87	1.00

3.3 Classifier

Deep learning frameworks have always prioritized either speed or usability over the other. PyTorch is a machine learning framework that demonstrates that these two objectives may coexist [30]. We utilized the Pytorch framework to create the model and for training purposes, among other things. In our work, we didn't use a pre-trained network, and we had to start from scratch with the training. On the basis ofAlexNet architecture, we've created a network. AlexNet is a large, deep CNN that uses 1000 different classes to classify the 1.3 million high-resolution images in the LSVRC-2010 ImageNet training set [31]. When compared to ResNet model topologies and other common networks, the proposed model is much lighter. The parameters in our model total roughly 50K. The network's architecture is seen in Fig.1 and 2. Fig. 3 depicts the model's whole architecture. Sequential Block is a torch.nn. Sequential object with i=l,2,3,4in Fig. 4 depicts a single convolution block with four layers: convolution, batch normalization, dropout and ReLU (Rectified linear activation unit) activation. SGD (Stochastic Gradient Descent) optimizer and OneCycleLR scheduling policy are used to train the proposed model. Super convergence is achieved by training the model with cyclical learning rates rather than constant values, which results in enhanced classification accuracy without the need to weak the hyper parameters and typically in less iteration [32].

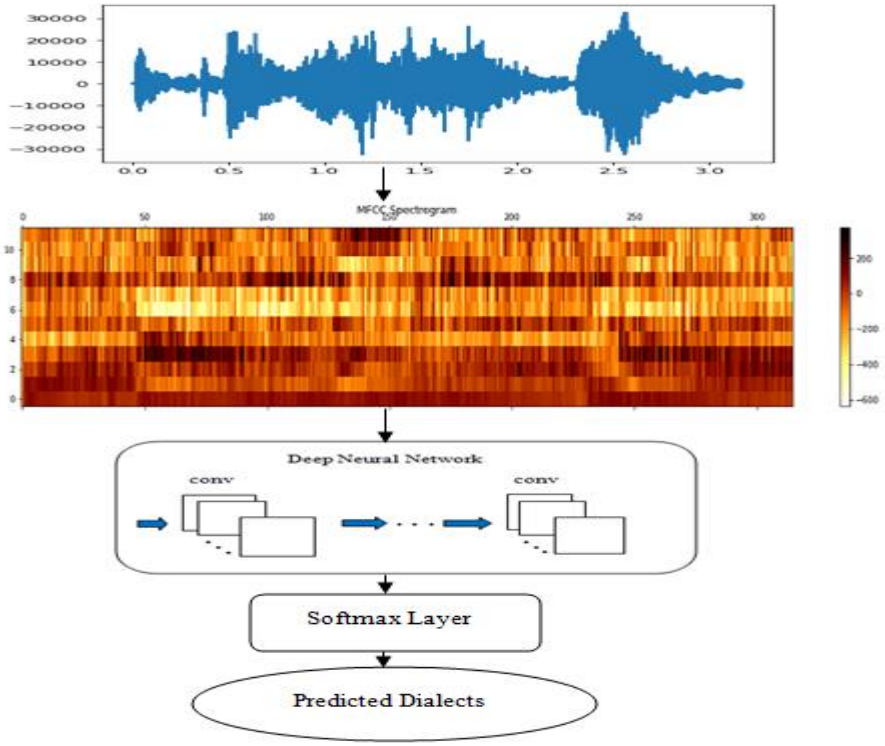

Figure 1. The CNN Model for Goalporia dialect audio file of duration 3 seconds.

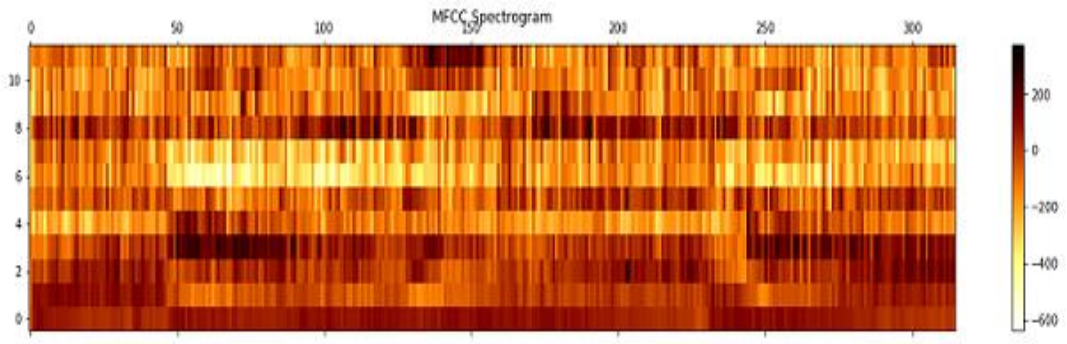

Figure 2. Mel spectrogram of Goalporia dialect audio file of duration 3 seconds.

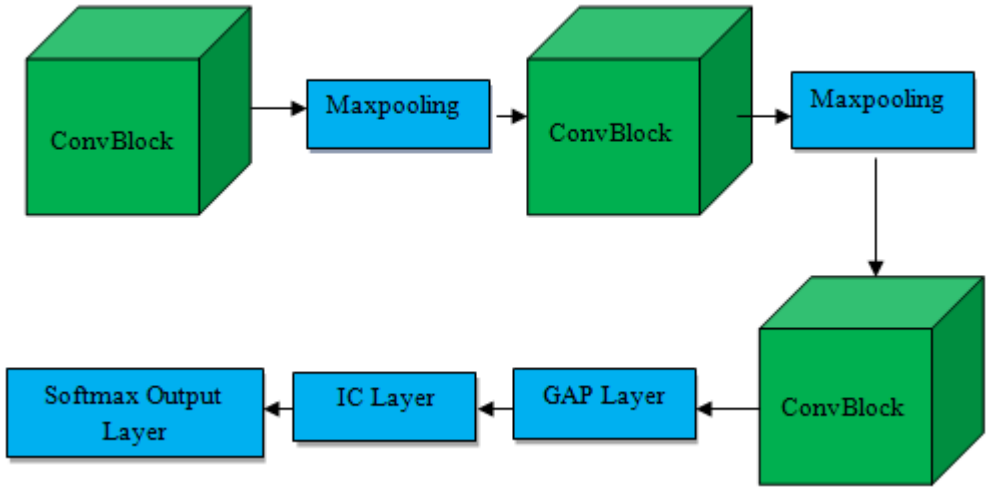

Figure 3. The architecture of the proposed model [33].

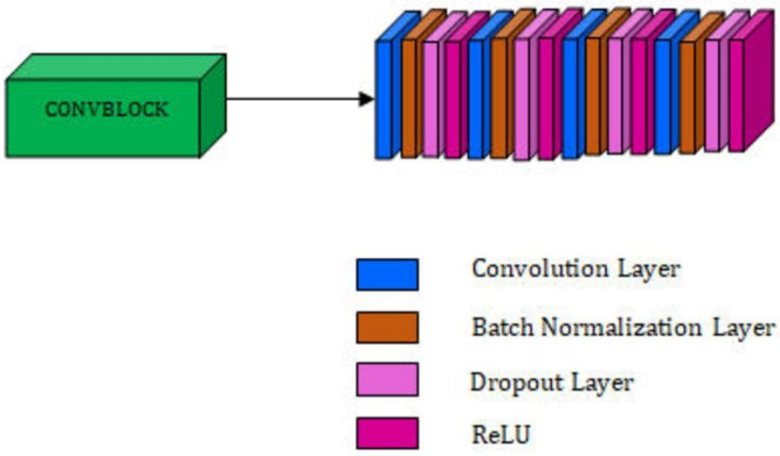

Figure 4. A convolution block in detail [33]

4.0 RESULTS AND DISCUSSION

Initially, we used a dataset of four dialects of Assamese languages to train our model. In training and testing, each audio clip of two seconds is used. The NLL loss (negative log likelihood loss) function was employed. The accuracy is calculated using the formula below.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ Number\ of\ predictions} \times 100 \qquad (2)$$

Fig. 3 shows the proposed model, which was trained by varying the learning rate and duration of the utterances taken from different dialects. OneCycleLR is the scheduling policy used, and it expects a maximum learning

rate as a hyper parameter. A minimal Learning Rate is derived based on the number of epochs and the remaining parameters. We tried three different audio lengths (2, 3, and 4 seconds) as well as two different maximum Learning Rates (0.0725 and 0.1). The relationship between audio duration and system accuracy is shown in Table 3. The results show that SLID's accuracy improves as the length of the audio file increases.

Table 3. Accuracy vs. audio sample size

Duration of Audio	Learning Rate	Accuracy
2	0.0725	79.46
2	0.1	78.39
3	0.0725	79.02
3	0.1	83.31
4	0.0725	80.4
4	0.1	83.52

A separate version of the model with 40 million parameters gives the best accuracy of 89.92 %.This finding demonstrates that model capacity is important for accuracy, but it also increases training time. We achieved an accuracy of 93% in predicting Eastern dialect, 87% in predicting Central dialect, 81% in predicting Kamrupia dialect, and 78 % in predicting Goalporia dialect in this experiment.
The confusion matrix in Table 4 shows that the majority of Eastern dialect, Central dialect, and Kamrupia dialect were properly predicted. However, in case of Goalporia the prediction accuracy is less.

Table 4. Confusion matrix

	Eastern dialect	Central dialect	Kamrupia dialect	Goalporia dialect
Eastern dialect	1151	186	76	0
Central dialect	62	444	22	0
Kamrupia dialect	2	163	475	31
Goalporia dialect	2	2	62	755

Hyperparmeters are variables that govern the learning process and have predetermined values. We need to fine-tune them by running numerous tests with various variables and selecting the best ones. Learning rate, momentum, and duration of each training or testing audio samples are some of the key factors we use in our work. We experimented with various learning rates and durations. The best results were obtained with a learning rate of 0.0275 and duration of 4 seconds for each training and testing audio samples.

5.0 CONCLUSION

In Deep Learning models built on CNNs, image data is used more frequently than text and voice data. Through the application of deep learning techniques to audio data, the CNN-based recommended model achieved a notable level of accuracy and performance. Stable performance is always aided by increased data, which we do in our work by employing audio changes like frequency and temporal masking. Our research led us to three conclusions. The accuracy of dialect identification is influenced by the duration of the audio. While heavier models outperform lighter models in terms of performance, they also take longer to train. Lastly, the bulk of eastern audio recordings are estimated to be Central due to the striking similarity between the sounds of the Eastern and Central dialects. Given the degree of similarity between regional dialects, we will need to utilize very powerful models in order to get reasonable performance. Using sequential models such as RNNs, LSTMs, GRUs, Bidirectional GRUs, and Transformers can help us improve the performance of our models. In most Spoken dialect identification applications, the source of audio data will not be controllable. Therefore, it's imperative to create models that perform well under a variety of noise distributions, speaker variability, distinctive accents, and speaker genders and age groups. As a future endeavor, we can experiment with adding noise and speeding up the audio to improve the model's accuracy.

REFERENCES

1. C.C.Fries, K.L.Pike., “Coexistent phonemic systems. Language.*JSTOR*, Vol. 25, No. 1 (Jan. - Mar., 1949), pp. 29-50.
2. Assamese Website, in Resource centre for indian language technology solutions RCILTS-II, IIT Guwahati 2006. http://www.iitg.ernet.in/rcilts/Assamese-language. Accessed 7 Dec 2016.
3. Wikipedia, Assamese language, 2019. https://en.wikipedia.org/wiki/Assamese language. Accessed 10 Oct 2019.
4. G.A.Liu, J.H.L. Hansen, “A systematic strategy for robust automatic dialect identification*”, in 2011 IEEE 19th European Signal Processing Conference (EUSIPCO 2011)*, Barcelona, Spain, 29 August 2011 - 02 September 2011.

5. H. Li, B. Ma, K.A. Lee, “Spoken language recognition: From fundamentals to practice”.*Proceedings of the IEEE*, Vol. 101, No. 5, May 2013, pp. 1136-1159. doi:10.1109/JPROC.2012.2237151.
6. J. Zhao, H. Shu, L. Zhang, X. Wang, Q. Gong, “Cortical Competition during Language Discrimination”.*NeuroImage*, Vol. 43, No.3,November 2008, pp. 624-633. doi:10.1016/j.neuroimage.2008.07.025.
7. P. Heracleous, K. Takai, K. Yasuda, Y. Mohammad, A. Yoneyama, “Comparative Study on Spoken Language Identification Based on Deep Learning”, *in 2018 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy,03-07 September 2018. doi:10.23919/EUSIPCO.2018.8553347.
8. I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D.Martinez, J.G. Rodriguez, “Automatic language identification using deep neural networks”, *in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 04-09 May 2014.doi:10.1109/ICASSP.2014.6854622.
9. G. Montavon, “Deep learning for spoken language identification”,*NIPS Workshop on deep learning for speech recognition and related applications*,2009.
10. A.A. Nti, “Studying dialects to understand human language, Dissertation”,*Massachusetts Institute of Technology*, Massachusetts, ME, 2009.
11. N,C,J, Bailey, “Is There a ”Midland” Dialect of American English?”. ERIC Clearinghouse, 1968.
12. L.M. Davis, C.L. Houck, “Is There a Midland Dialect Area?—Again”.*American Speech*, Vol. 67, No. 1, 1992, pp. 61-70. doi: 10.2307/455758.
13. A. Etman, A.L.Beex, “Language and Dialect Identification: A survey”, *in 2015 SAI Intelligent Systems Conference (IntelliSys)*, London, UK, 10-11 November 2015, pp.220-231. doi:10.1109/IntelliSys.2015.7361147.
14. M. Diab, N. Habash, “Arabic dialect processing tutorial”, *in Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, Rochester, 2007, pp.5-6.
15. J. Watson, *The Semitic Languages: An international handbook*, Weninger, S., Khan, G., Streck, M., Watson, J.C.E. (eds.) ,Berlin,Walter de Gruyter, 2012,Arabic dialects (general article),pp 851-896.
16. N.J. Ibrahim, M.Y.I. Idris, M. Yakub, N.N.A. Rahman, M.I. Dien, “Robust feature extraction based on spectral and prosodic features for classical Arabic accents recognition”.*Malaysian Journal of Computer Science*, No.3, pp. 46-72. doi:10.22452/mjcs.sp2019no3.4.
17. S. Shivaprasad, M. Sadanandam, “Identification of regional dialects of Telugu language using text independent speech processing models”.*International Journal of Speech Technology*, Vol. 23, February 2020, pp. 251-258.https://doi.org/10.1007/s10772-020-09678-y.
18. N.B. Chittaragi, S.G. Koolagudi, “Acoustic features based word level dialect classification using SVM and ensemble methods”, *in 2017 Tenth International Conference on Contemporary Computing (IC3)*, Noida, India,10-12 August 2017, pp. 1–6.
19. N.B. Chittaragi, A. Limaye, N. Chandana, B. Annappa, S.G. Koolagudi, “Automatic text-independent kannada dialect identifcation system”,*in Proceedings of Information Systems Design and Intelligent Applications*, 2019, pp. 79–87.
20. K.S. Rao, S.G. Koolagudi, “Identification of Hindi dialects and emotions using spectral and prosodic features of speech”.*IJSCI: International Journal of Systemics, Cybernetics and Informatics*, Vol. 9, No. 4, 2011, pp. 24–33.
21. B. Aarti, S.K. Kopparapu, “Spoken Indian language classification using artificial neural network — An experimental study”,*in 2017 4th International Conference on signal processing and integrated networks (SPIN)*, Noida, India, 2017, pp. 424–430.
22. C. Madhu, A. George, L. Mary, “Automatic language identification for seven indian languages using higher level features”,*in 2017 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, Kollam, India, 2017, pp. 1–6.
23. M. K. Veera, R.K. Vuddagiri, S.V. Gangashetty, A.K. Vuppala, “Combining evidences from excitation source and vocal tract system features for Indian language identification using deep neural networks”. *International Journal of Speech Technology*, Vol. 21, No.3, December 2017, pp. 501–508.
24. I. Lopez-Moreno et al., “Automatic language identification using deep neural networks”,*in 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Florence, Italy, 2014, pp. 5337–5341.
25. J. Boussard, A. Deveau, J. Pyron, “Methods for spoken language identification”,Stanford University, 2017.
26. C. Ittichaichareon, S. Suksri, T. Yingthawornsuk, “Speech recognition using mfcc”, *in International conference on computer graphics, simulation and modeling*, Florence, Italy ,04-09 May 2014, 2012.
27. N. Sharma, V. Jain, A. Mishra, “An analysis of convolutional neural networks for image classification”,*Procedia computer science,* Vol. 132, 2018,pp. 377–384.
28. torchaudio−Torchaudio master documentation(2020). Pytorch.org. Available: http://pytorch.org/audio/. Accessed 9 Sept 2020.
29. Purwins, B. Li, T. Virtanen, J. Schlüter, S.Y. Chang, T. Sainath, “Deep learning for audio signal processing”, *IEEE Journal of Selected Topics in Signal Processing* ,Vol.13, No.2, 2019, pp.206–219.
30. A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library”.*Advanced in Neural Inf Processing Systems (NeurIPS),* Vol. 32, 2019, pp. 1-12.

31. A. Krizhevsky, I. Sutskever, G.E. Hinton, “Imagenet classification with deep convolutional neural networks”, *Communications of the ACM* , Vol. 60, No.6, 2017, pp. 84–90.
32. L.N. Smith, “Cyclical learning rates for training neural networks”, *in 2017 IEEE winter conference on applications of computer vision (WACV)*, Santa Rosa, CA, USA , 2017, pp. 464–472.
33. L.R. Arla, S. Bonthu, A. Dayal, “Multiclass spoken language identification for Indian languages using deep learning”, *in 2020 IEEE Bombay Section Signature Conference (IBSSC)*, Mumbai, India, 2020,pp. 42-45.

DOI: https://doi.org/10.15379/ijmst.v10i2.3519