Assamese Dialect Identification System Using Deep Learning

Hem Chandra Das (hemchandradas78@gmail.com)
Bodoland University
Prof.utpal Bhattacharjee
Rajiv Gandhi University

Research Article

Keywords: CNN, DID, Spectrogram, Classification, Assamese, Dialect Identification, Deep Learning

Posted Date: June 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1733629/v1

Assamese Dialect Identification System Using Deep Learning

Hem Chandra Das[1,2*] and Utpal Bhattacharjee[2,1]

[1*]Computer Science and technolog, Bodoland University, Kokrajhar,, 783370, Assam, India.
[2]Computer Science and Engineerin, Rajiv Gandhi University, Papum Pare, 791112, Arunachal Prades, India.

*Corresponding author(s). E-mail(s): hemchandradas78@gmail.com;
Contributing authors: utpal.bhattacharjee@rgu.ac.in;

Abstract

The goal of a dialect Identification is to label speech in an audio file with dialect labels. This paper presents a method for automatically identifying four Assamese dialects: Central Assamese, Eastern Assamese dialect, Kamrupi dialect, and Goaplari dialect, using Convolution Neural Networks (CNN). In this study, utterances of four major regional dialects of the Assamese language, namely Central Assamese spoken in and around Nagaon district, Eastern Assamese dialect spoken in the Sibsagar and its neighboring districts, Kamrupi dialect spoken in Kamrup, Nalbari, Barpeta, Kokarajhar and some parts of Bongaigaon district and Goaplari dialect spoken in the Goaplara, Dhuburi and part of Bongaigaon district were used. The classifier was trained on audio samples from each of the four dialects that lasted 2 hours. The CNN uses Mel spectrogram images created from two to four seconds divisions of raw audio input with varied audio quality. The performance of the system is also examined as a function of train and test audio sample durations. When compared to machine learning models, the suggested CNN model obtains an accuracy of 90.82 percent, which may be considered the best.

Keywords: CNN, DID, Spectrogram, Classification, Assamese, Dialect Identification, Deep Learning

1 Introduction

At present, Dialect identification has an important research topic of signal processing because it has a wide range of applications in our daily lives. The identification of languages from spoken utterances is known as Dialect identification. One of the difficulties in this work is properly and quickly identifying dialect from audio at varying sample speeds and noise levels. Another difficulty is distinguishing between dilects that sound similar over a short period of time [1]. Assamese language is derived from the Indo-Aryan family of languages and is spoken by the majority of the natives of Assam, a state in North East India and in some parts of its neighboring states like Meghalaya, Nagaland and Arunachal Pradesh. Assamese is the official language of the state of Assam and it is listed in schedule-VIII of the Indian Constitution as a Major Indian Language. The Assamese language evolved from Sanskrit, but the original inhabitants of Assam, such as the Bodos and Kacharis, had a great influence on its vocabulary, phonology, and grammar [2]. According to recent research, there are four major dialect groups of the Assamese language. The Eastern group is spoken in the Sibsagar district and its surrounding areas, while the Central group is spoken in and around the present-day Nagaon district and its surrounding areas. The Kamrupi dialect is spoken in unincorporated Kamrup, Nalbari, Barpeta, Darrang, and a portion of Bongaigaon. Goalpara, Dhubri, and portions of Kokrajhar and Bongaigaon districts are home to the Goalparia group . However, presently, the Central Assamese is widely considered as the primary or standard dialect [2]. In the field of Assamese dialect translation, there is no significant has been reported. In linguistics, a dialect is a type of language that is socially distinct and is spoken by a specific group of native speakers who have a similar pattern of pronunciation, syntax, and vocabulary [3]. The ability to differentiate between spoken languages is a feature of the human intelligence [4] [5]. The first stage in developing a dialect-independent voice recognition system for any language is dialect identification. In recent decades, the CNN models have yielded promising results [6]. To complete this objective, several deep neural network models were investigated, as it has already been reported that such architectures give good results [7] [8]. These families of models are able to extract significant features automatically; however they all require a preprocessing step of audio to frequency domain translation. In this work deep learning techniques such as convolutions and maxpoolings have been used to identify Assamese dialect. Because we used a modest model (a Convolutional neural network with fewer parameters), training and testing are simple, and training takes very little time. Our model converts a batch of raw audio data into a batch of mel spectograms for training or validation purposes.

2 Literature Review

George Wenker undertook a series of studies to determine dialect regions in 1877, which started the field of dialect identification [9]. Baily was one of

the pioneers in the identification and establishment of the Midland dialect as a distinct dialect. Following the findings of the study, it was concluded that dialects should not be classified only on the basis of vocabulary, because vocabulary might vary significantly amongst groups or classes within a particular geographic area [10]. Davis and Houck [11] also attempted to determine whether the Midland region can be considered as a separate dialect region. The researchers effectively extracted phonological and lexical characteristics from 11 cities along the north-south line [12]. Diab et al. [13] and Watson [14] analyzed the Arabic dialect, listed its characteristics, determined the link between the Standard language and its regional variations, and classified the major regional dialects. Ibrahim et al. [15] use GMM to identify Arabic dialects using Spectral and prosodic characteristics. In Malaysian Quranic Dialect recognition 5.5 to 7Dialect recognition researches have been done in many Indian languages. Shivaprasad and Sadanandam [16] used GMM and HMM to identify regional Telugu dialects. The authors generated a database of Telugu dialects for this purpose. Recogntion was done using MFCC and its variants such as ΔMFCC and ΔΔMFCC features. The study extracts 39 feature vectors from each spoken utterance and evaluates them with GMM and HMM models. The GMM model outperforms the HMM model. However, certain words with identical acoustic characteristics were not distinguished. Chittaragi and Koolagudi [17] use the closest neighbor approach to identify Telugu dialects using just prosodic characteristics and a few lines from each dialect. Authors attained 75Chittaragi et al. [18]] discovered 5 Kannada dialects using spectral and prosodic characteristics. The authors used SVM and Neural Networks to detect dialects. The Neural Network gives good outcomes with text-independent data in the quickest time. K. S. Rao et al. [19] utilized spectral and prosodic characteristics to distinguish five Hindi dialects: Chattisgharhi, Bengali, Marathi, General and Telugu. Their database comprises ten (10) individuals speaking spontaneously for 5-10 minutes each, totaling 1-1.5 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 [20]. 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-ΔMFCC, and 13-ΔΔMFCC 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. [21] used an ANN classifier to identify seven Indian languages, including Bengali, Telugu, Urdu, Hindi, Assamese, Manipuri, and Punjabi, using a 2-hour voice database. They attained accuracy of 72Veera et al. [22] compared Language Identification system utilising two methods: Deep Neural Network with Attention (DNN-WA)

and i-vector system. Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Manipuri, Marathi, Odiya, Punjabi, Tamil, Telugu, and Urdu were among the 13 Indian languages represented in the sample, which included 30 hours of test data and 15 hours of training data. Shifted Delta Cepstral coefficients (SDC) were applied to both MFCC and Residual Cepstral Coefficients (RCC) to create a LID system. Using RCC and MFCC features, respectively, EERs of 9.93% and 6.25% were achieved. Using a method known as late fusion, an EER of 5.76 percent was also obtained by fusing the characteristics from both features. 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. suggested that using short temporal acoustic characteristics, a deep neural architecture can determine the language of a spoken speech [23]. The NIST Language Recognition Evaluation (LRE) 2009 and the Google 5M LID corpus were both used in their study. The performance of this DNN-based acoustic model was compared to that of an i-vector model. The DNN model outperformed the i-vector-based technique when dealing with large amounts of data. Using the OGI Multilanguage Corpus, Boussard et al.[24] studied LID using phone calls dataset of eleven languages having a duration ranging from a few seconds to almost a minute. To create predictions directly on the call level or from the derived features such as MFCC, SDC and spectral centroids, many classification techniques were utilized, including aggregating data at the frame level, Feed-Forward Neural Networks, Recurrent Neural Networks (RNN), CNN, and Gaussian Mixture Models (GMM). The results revealed that GMM, when combined with SDC features, delivered the best outcomes.

3 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.

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

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

mel-cepstrum makes use of auditory concepts as well as the cepstrum's decorrelating characteristic [25]. 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.

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

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 [26]. As a result, we trained our classifier using Deep Learning techniques such as Convolutions and Maxpoolings. We used two to four second audio tracks for the training procedure. Because our algorithm was trained on two-second audio files, it only requires a two to four second audio clip to predict the dialect. As a result, the dialect detection procedure is extremely quick. Figure 1 depicts the detailed architecture. The raw audio signals are converted to Mel spectrograms, allowing it to escape over fitting. To get a dialect identification, these spectrograms are fed into our suggested model, which is based on AlexNet. 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 organise our data in a logical manner, we utilised 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 speakers are of various genders and have different accents. Each audio file is transformed to melspectograms during training. Melspectograms were

created with Pytorch's torchaudio library [27]. 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. Time masking and frequency masking are the transformations used here. Frequency masking refers to the application of masking to a spectogram in the frequency domain, while time masking refers to the application of masking to a spectogram in the time domain [28]. 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 [29]. 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 of AlexNet 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 [30]. 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 Figures 3 and 4. Figure 3 depicts the model's whole architecture. Sequential Blocki is a torch.nn. Sequential object with i=l,2,3,4. Figure 4 depicts a single convolution block with three layers: convolution, batch normalization, 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 tweak the hyper parameters and typically in less iteration [31].

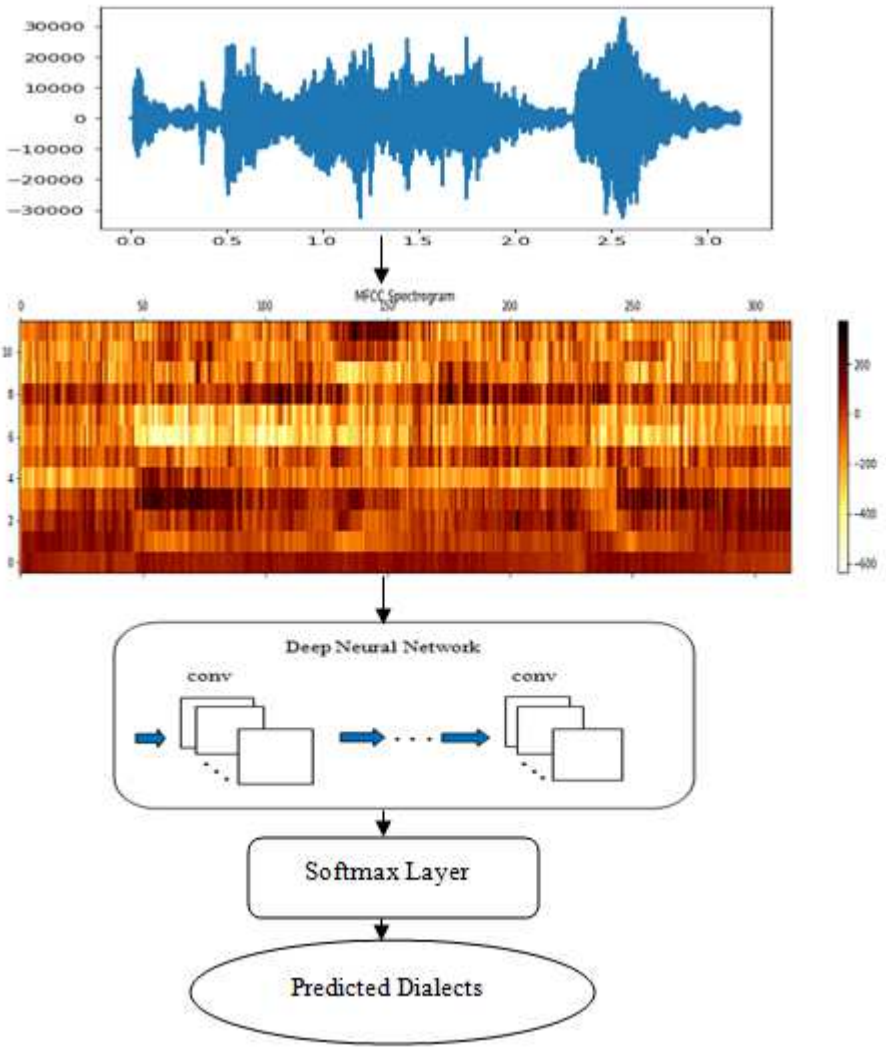

Fig. 1 Mel spectrogram of an audio file of duration 3 seconds

4 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 function was employed. The accuracy is calculated using the formula below.

$$Accuracy = \frac{No of correct predictions}{Total number of predictions} \times 100. \quad (2)$$

Figure 2 shows the proposed model, which was trained by varying the learning rate and duration of the utterances taken from different dialects. OneCycleLR

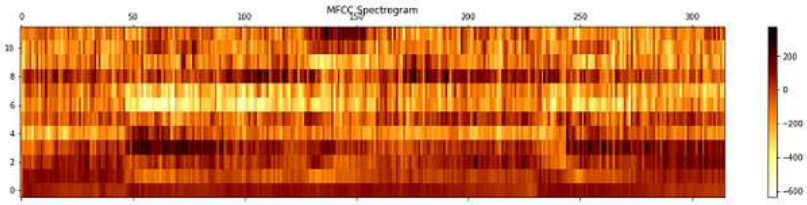

Fig. 2 Mel spectrogram of an audio file of duration 3 seconds

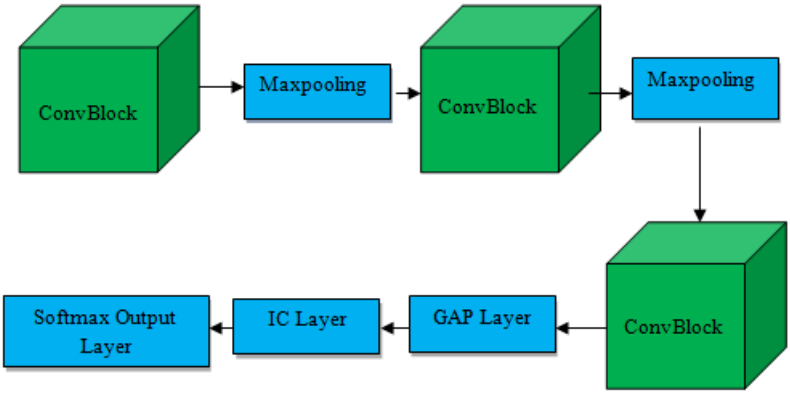

Fig. 3 The architecture of the proposed model[32]

is the scheduling policy used, and it expects a maximum learning rate as a hyperparameter. 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. A separate version of the model with 40 mil-

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

lion 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

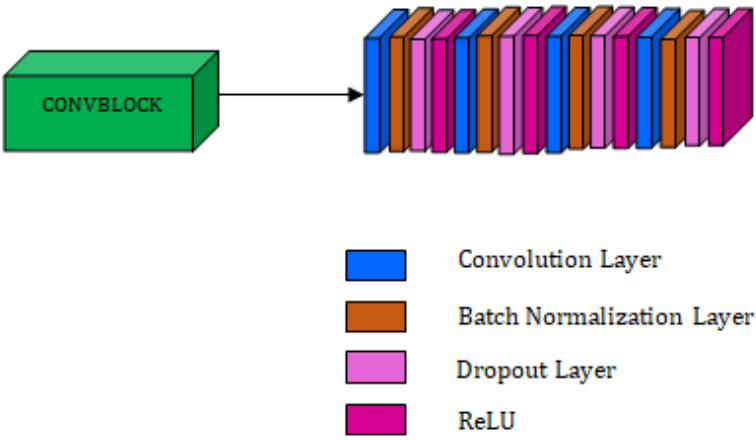

Fig. 4 A convolution block in detail[32]

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. Hyperparmeters are variables

Table 4 Confusion matrix

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

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 Conclusion

Image data is more commonly employed than text and voice in Deep Learning models based on CNNs. By applying deep learning techniques to audio data, the suggested model based on CNNs acquired a noteworthy accuracy

and good performance. Increased data always aids in achieving stable performance, which we do in our work by using audio modifications such as temporal and frequency masking. We made three observations as a result of our work. The length of the audio has an impact on the dialect identification accuracy. Although heavier models perform better than lighter models, they require more training time. Finally, because Eastern and Central dialect sounds are so similar, the majority of eastern audio files are projected to be Central. We need to employ some strong models to obtain decent performance because there is a lot of a similarity in the regional dialects. We can increase the performance of our models by using sequential models like as RNNs, LSTMs, GRUs, Bidirectional GRUs, and Transformers. There will be no control over the source of audio data in most Spoken dialect Identification applications. As a result, it's critical to develop models that operate effectively when subjected to varied noise distributions, speaker's variability, distinct accents, and different genders and age groups of speakers. As a future endeavor, we can experiment with adding noise and speeding up the audio to improve the model's accuracy.

6 Declarations

6.1 Ethical Approval and Consent to participate

I, the undersigned, offer my approval for identifiable details to be published in the aforesaid Journal and Article, which may include photograph(s), videos, case history, and/or details within the text ("Material").

6.2 Consent for publication

Not Applicable

6.3 Availability of supporting data

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. Therefore, Raw data were generated by 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. large-scale facility. Derived data supporting the findings of this study are available from the corresponding author upon request.Derived data supporting the findings of this study are available from the corresponding author upon request.

6.4 Competing interests

Not Applicable

6.5 Funding

Not Applicable

6.6 Authors' contributions

Conceptualization: Hem Chandra Das, Utpal Bhattacharjee, Methodology: Hem Chandra Das, Utpal Bhattacharjee,Formal analysis and investigation: Hem Chandra Das, Utpal Bhattacharjee, Writing - original draft preparation: Hem Chandra Das; Writing - review and editing: Hem Chandra Das, Utpal Bhattacharjee, Funding acquisition: [Not applicable], Resources: [Hem Chandra Das],Supervision: Utpal Bhattacharjee

6.7 Acknowledgments

Not applicable

References

[1] Pi School - Machine Intelligence meets Human Creativity.Spoken Language Identification - Pi School- Machine Intelligence meets Human Creativity.https://picampus-school.com/spoken-language-identification(2018).Accessed 15 Sept 2020

[2] Wikipedia(2019).Assamese language.https://en.wikipedia.org/wiki/Assamese language.Accessed 10 oct 2019

[3] Liu, G.A.,Hansen,J.H.L.:A systematic strategy for robust automatic dialect identification.In:2011 IEEE 19th European Signal Processing Conference (EUSIPCO 2011), Barcelona, Spain; 2011. pp. 2138-2141.IEEE,Spain(2011)

[4] Li, H.,Ma, B.,Lee, K.A.:Spoken language recognition: From fundamentals to practice.Proceedings of the IEEE 2013,101(5),1136-1159(2013). doi:10.1109/JPROC.2012.2237151

[5] Zhao, J., Shu, H., Zhang, L., Wang, X., Gong, Q.: Cortical Competition during Language Discrimination.NeuroImage. 43(3), 624-633(2008). doi:10.1016/j.neuroimage.2008.07.025

[6] Heracleous, P., Takai, K., Yasuda, K., Mohammad, Y., Yoneyama, A.:Comparative Study on Spoken Language Identification Based on Deep Learning.In:2018 26th European Signal Processing Conference

(EUSIPCO), Rome, Italy, September 2018. IEEE xplore, pp. 2265-2269.IEEE,Italy(2018)

[7] Lopez-Moreno,I., Gonzalez-Dominguez, J., Plchot, O., Martinez, D., Rodriguez, J.G.:Automatic language identification using deep neural networks.In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),Florence, May 2014. IEEE Xplore, pp.5337-5341.IEEE,Italy(2014)

[8] Montavon, G.: Deep learning for spoken language identification. NIPS Workshop on deep learning for speech recognition and related applications, Citeseer, pp. 1-4(2009)

[9] Nti, A.A.: Studying dialects to understand human language. Dissertation, Massachusetts Institute of Technology (2009)

[10] Bailey, N.C.J.: Is There a "Midland" Dialect of American English?. ERIC Clearinghouse(1968)

[11] Davis L.M., Houck, C.L.: Is There a Midland Dialect Area?—Again:American Speech.67(1):61-70(1992)

[12] Etman, A., Beex, A.L.:Language and Dialect Identification: A survey. In: 2015 SAI Intelligent Systems Conference (IntelliSys), London, UK, 2015.IEEE Explore,pp.220-231.IEEE(2015)

[13] Diab, M., Habash, M.:Arabic dialect processing tutorial.In:Proceedings of the Human Language Technology Conference of the NAACL, Rochester,New York, USA,2007.Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, Tutorial Abstractspp, pp.5-6

[14] Watson, J.:Arabic dialects (general article). In: In: Weninger, S., Khan, G., Streck, M., Watson, J.C.E. (eds.) The Semitic Languages: An international handbook,pp. 851-896. Walter de Gruyter, Berlin(2011)

[15] Ibrahim, N.J.,Idris M.Y.I.,Yakub, M.,Rahman, N.N.A.,Dien, M.I.:Robust feature extraction based on spectral and prosodic features for classical Arabic accents recognition.Malaysian Journal of Computer Science.3,46-72(2019). doi:10.22452/mjcs.sp2019no3.4

[16] Shivaprasad, S., Sadanandam, M.:Identification of regional dialects of Telugu language using text independent speech processing models.International Journal of Speech Technology.23, 251-258(2020).doi: 10.17485/ijst/2017/v10i20/115033

[17] Chittaragi, N.B., Koolagudi, S.G.:Acoustic features based word level dialect classification using SVM and ensemble methods. In:2017 Tenth International Conference on Contemporary Computing (IC3), Noida, August 2017, pp. 1-6.IEEE Xplore,India(2018)

[18] Chittaragi, N.B., Limaye, A., Chandana, N.P., Annappa, B., Koolagudi, S.G.:Automatic text-independent Kannada dialect identification system.Information Systems Design and Intelligent Applications,863, 79-87(2019). doi:10.1007/978-981-13-3338-5_8

[19] Rao, K.S., Koolagudi, S.G.:Identification of Hindi dialects and emotions using spectral and prosodic features of speech.International Journal of Systemics, Cybernetics and Informatics. 9(4), 24-33(2011)

[20] Aarti, B., Kopparapu, S.K.:Spoken Indian language classification using artificial neural network — An experimental study. In:2017 4th International Conference on Signal Processing and Integrated Networks (SPIN);Noida,February 2017.IEEE Explore,pp.424-430.IEEE,India(2017)

[21] Madhu, C., George, A., Mary, L.: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,August 2017. IEEE Explore,pp.1-6.IEEE,India(2017)

[22] Veera, M.K., Vuddagiri, R.K.,Gangashetty, S.V.,Vuppala, A.K.:Combining evidences from excitation source and vocal tract system features for Indian language identification using deep neural networks.International Journal of Speech Technology.21,501-508(2017). doi: 10.1007/s10772-017-9481-6

[23] Lopez-Moreno, I.,Gonzalez-Dominguez, J.,Plchot, C.,Martinez, D., Gonzalez-Rodriguez, J.,Moreno, P.:Automatic language identification using deep neural networks.In:2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence,May 2014.IEEE Explore, pp.5337-5341.IEEE,Italy(2014)

[24] Boussard, J., Deveau, A., Pyron, J.:Methods for Spoken Language Identification.Stanford University(2017)

[25] Ittichaichareon, C., Suksri, S., Yingthawornsuk, T.:Speech Recognition using MFCC.In:International conference on computer graphics, simulation and modeling, Pattaya,Thailand, 28-29 July 2012

[26] Sharma, N., Jain, V., Mishra, A.:An Analysis Of Convolutional Neural Networks For Image Classification.Procedia computer science.132, 377-384(2018).doi: 10.1016/j.procs.2018.05.198

[27] torchaudio−Torchaudio master documentation.Pytorch.org.http://pytorch.org/audio(2020).Accessed 9 Sep 2020

[28] Purwins, H., Li, B., Virtanen, T.,Schlüter, J.,Chang, S.,Sainath, T.:Deep Learning for Audio Signal Processing.IEEE Journal of Selected Topics in Signal Processing.13,206-219(2019).doi:10.1109/JSTSP.2019.2908700

[29] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., et al.:Pytorch: An imperative style, high-performance deep learning library.Advanced in Neural Information Processing Systems (NeurIPS 2019).32,1-12(2019)

[30] Krizhevsky, A., Sutskever, I., Hinton, G.E.:ImageNet Classification with Deep Convolutional Neural Networks.Advances in neural information processing systems.25,1-9(2012)

[31] Smith, L.N.:Cyclical Learning Rates for Training Neural Networks.In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV),Santa Rosa, CA, March 2017. IEEE Explore,pp. 464-472.IEEE,USA(2017)

[32] Arla, L.R.,Bonthu, S.,Dayal, A.:Multiclass Spoken Language Identification for Indian Languages using Deep Learning.In:2020 IEEE Bombay Section Signature Conference (IBSSC),Mumbai,December 2020.IEEE Explore,pp. 42-45,IEEE,India(2020)