International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2390556

Utilizing Deep Learning Techniques for the Classification of Spoken Languages in India

Priyesha Patel*, Ayushi Falke, Dipen Waghela, Shah Vishwa

Computer Engineering, Parul University, Post Limda, Waghodia, Gujarat, India

ARTICLEINFO

Article History:

Accepted: 01 March 2024

Published: 11 March 2024

Publication Issue

Volume 10, Issue 2

March-April-2024

Page Number

63-69

ABSTRACT

In Western countries, speech-recognition applications are accepted. In East Asia, it isn't as common. The complexity of the language might be one of the main reasons for this latency. Furthermore, multilingual nations such as India must be considered in order to achieve language recognition (words and phrases) utilizing speech signals. In the last decade, experts have been clamoring for more study on speech. In the initial part of the pre-processing step, a pitch and audio feature extraction technique were used, followed by a deep learning classification method, to properly identify the spoken language. Various feature extraction approaches will be discussed in this review, along with their advantages and disadvantages. Also discussed were the distinctions between various machine learning and deep learning approaches. Finally, it will point the way for future study in Indian spoken language recognition, as well as AI technology.

Keywords: Speech Recognition, Indian Language, Spoken Language, Pitch, Audio Feature, Machine Learning and Deep Learning

I. INTRODUCTION

Language identification refers to a machine's capacity to recognise spoken language. Detection of spoken language is done automatically using language recognition. Speeches are usually given by a stranger. The use of voice command systems to link humans and machines is becoming more common in today's world. Those who are fluent in the spoken language may now be confidently identified.

Consequently, persons in South Asian nations have been unable to fully benefit from advancements in speech recognition technology since the development of speech detection algorithms for the Indic languages has been delayed as a result of this delay. Because of their complexity, Indian languages are difficult to convey on their own, and the multilingualism of these nations makes the task much more difficult. When speaking in this country, it is vital to identify the language of uttered words and phrases before attempting to recognise them since individuals seldom communicate in a combination of languages. Automated speech authentication is a technique for automatically distinguishing between different

languages based on voice cues. This system can distinguish spoken segments and activate language-specific recognizers, which is useful in multilingual nations such as India, where speech recognition is difficult.

There are two primary approaches to speech recognition: acoustic and phonetic. Initially, acoustic approaches recover short-term speech spectrum features as a multidimensional vector. A statistical model is built for each language based on the extracted features. In acoustic-based SLR systems, Gaussian mixture models are the most often used model (GMM). The bulk of acoustic-based language recognition systems today employ the i-vector technique. It's the best in the field of language detection. This technique converts each speech file into a fixed-length vector. i-vectors are compressed speech signals used as input feature vectors in recognition systems' classification steps. Short-term acoustic characteristics are the easiest way to extract information from a speech stream. It is possible to extract higher-level speech information, such as phonetic information, from the voice signal. Phonetic-based SLR systems utilise the speech signal's phonetic information.

In the next paragraphs, we'll discuss each of the following topics: In Section II, we'll look at some important advancements in speech recognition. Here, in Section III, a full description of the many approaches employed to create this framework is provided. Section IV is dedicated to comparative studies and discussions. Finally, some suggestions for further research are made at the conclusion of this study.

II. LITERATURE STUDY

TABLE I

LITERATURE PAPERS

Paper No.	Method	Limitation & Future Work
B. Paul et.al [1]	DNN, MFCC, LSTM, RNN, PMU.	To reduce the recognition time, increase the out of set database performance
H. S. Lee et.al [2]	RNN, DNN	In the development of PLLR, frame-based statistics or information are used.
M. A. A. Albadr et.al [3]	MFCC, SDC, GMM	The result demonstrated the superiority of the hypothesis. Because offline LID is considered ineffective, an online LID system is suggested to handle a larger variety of LID usage, including conferences and phone services.
H. Mukherjee et.al [4]	MFCC, SVM, CNN	Extremely competitive Make use of innovative approaches that open up a whole new world of possibilities.
D. S. Sisodia et.al [5]	GMM, SVM, HMM, MFCC	With the most up-to-date acoustic SLR technologies, Method systems are very competitive the precision and speed of the combined SLR systems are being increased.
D. S. Sisodia et.al [6]	MFCC, DFCC	Audio dataset recorded by the same speaker was the source of the problems. based on generic audio recordings recorded by various speakers with diverse tones and accents, the results were assessed
G. Singh et.al [7]	CNN, FFT	Utilize Neural Networks to improve performance.
H. S. Das et.al [8]	CNN, GMM, HMM, ANN	Working in real time is a difficult task. End-to-end DNNs that are hybrid in nature.
N. E. Safitri et.al [10]	CNN, SVM	A large dataset is not functioning properly. In the future, they will experiment with GMM and SVM on larger datasets.

P. Heracleous et.al [11]	NN, MFCC	Very Complex Feature with Less Accuracy
R. Fér et.al [12]	HMM, MFCC	Common Feature still need to use deep learning Concept.
O. Giwa et.al [13]	SVM	Exiting methods are occurring an issue then they solve in future.
R. W. M. Ng et.al [14]	MFCC, DNN	Use lower quality DNN. Improving the quality of the adapted DNN,
M. A. A et.al [15]	GMM	In this work, the LID System is used in an offline environment. It is necessary to design a LID system that can support online executions.
Y. Ma et.al [16]	AlexNet, Inception V3	Not work with Indian language of different region.
P. Beckmann et.al [17]	VGG Net	Spoken sentences do not recognize only words work.

III. PROPOSED METHODOLOGY

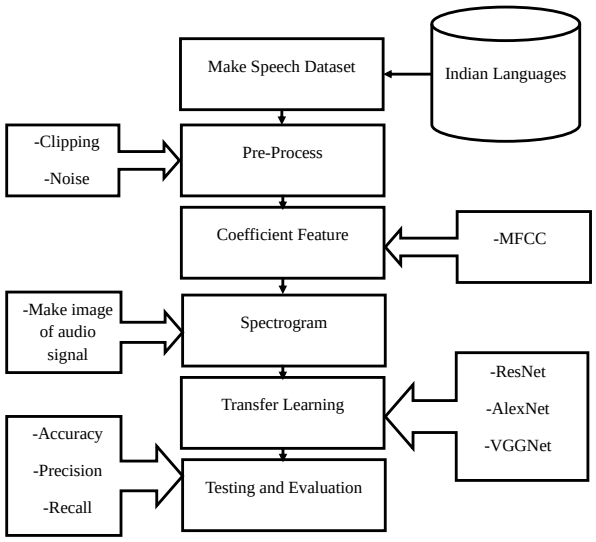

Figure 1. Proposed System

A. Datasets

The datasets in [1,4,6] are derived from the Indic speech corpus of the International Institute of Information Technology, Hyderabad (IIIT-H), which includes 1000 spoken sentences in each of seven different languages. Thus, they have utilised a total of 7000 audio samples in our language detection model.

There is a link to the data in [3] that may be found at (https://doi.org/10.6084/m9.figshare.6015173.v1).Additional data may be found on the author's website (http://www.ftsm.ukm.my/sabrina/resource.html), where it can be accessed.

The goal of this is to categories eight different languages, including English, Arabic, Malay, Spanish, French, German, Urdu, and Persian, into eight distinct categories. Each language has 15 utterances, with each speech lasting 30 seconds.

There are both target and non-target languages in the dataset set in [5]. With the use of this tool, SRL systems may be configured to identify Arabic, English, and Farsi languages by altering combination parameters or determining EER-based thresholds at operating locations across the system. Each target language has 200 files, while the non- target set has about 1000 files. The development files are typically about 30 seconds in duration.

The Kaggle dataset "spoken language identification" was used to analyses [7] distinct audio samples. These files include 10 second utterances, which are broken up into separate files.

B. Pre-Process [1,3,5,6]

In the second step of audio processing, known as "clipping," audio signals are broken into frequency frames that are of the same size. Once this is done, use the windowing feature to remove the borders. An energy spectrum with no crossovers at zero crossing rate. Remove background noise and unspoken information from an audio clip. Listeners should expect to hear 30 seconds of emotional variance in each track. Based on this evaluation, the start and finish points are determined.

Additional refinement of log-Mel spectrograms may be achieved by the removal of background noise from audio recordings. You may enhance the data by utilizing numerous techniques, such as pitch shifting and cropping and rotating and flipping as well as adding random noise and adjusting the audio speed.

C. Feature Extraction

There are three distinct spectrograms: [4,16,17]. Frequencies are found in varying amounts in real-world sounds, and the frequency components make up a sound signal's overall tone. What they hear is determined by the ratio of components and their frequencies. The spectral envelope is a term for this collection. The sound they perceive is heavily influenced by the form of this envelope. This envelope may be shown using spectrograms. It is made out of three-dimensional data. Sound clips are represented on the x-axis by their duration and the y- axis by their frequencies (the range of frequencies present). In the third dimension, color codes represent the energy of a specific frequency component at a given moment. Spectrograms may be shown in either grayscale or RGB. Audio signals are broken down into little chunks or frames, each of which undergoes the Fourier transformation in order to produce the spectrogram. The name "short-term Fourier transform" refers to this process.

TABLE II

CLASSIFICATION METHODS

Method	Advantage	Limitation
Support Vector Machine [1,4]	SVM is a simpler algorithm. Make very accurate classifiers. Less over-fitting, more noise resistant.	SVM is a binary classifier; pair-wise classifications may be 5 utilize 5 to perform multi-class classifications. Because it is computationally costly, it is sluggish.
Decision tree [11]	During Pre-preprocessing, less work is required for data preparation. Data does not need to be 5 utilize 5 ed or scaled. In addition, missing values in the data have little impact on the decision tree-building process. A decision tree paradigm is simple to understand and implement.	A slight change in the data can result in a substantial change in the decision tree's structure, resulting in instability. It takes longer to train the model. Because of the intricacy and time required, it is rather costly. Insufficient for predicting continuous values and using regression.
K-Nearest Neighbor [12]	Training data that is resistant to noise If the training data is substantial, it is effective.	When it comes to distance learning, it's not always apparent the sort of distance to 5 utilize or which characteristic to employ to get the greatest results.
Random Forest [6,8]	Improves the accuracy. Reduce over fitting. It can be used to solve both classification as well as regression problems.	Longer Training Period Due to Complexity
ResNet [13,16,17]	It is possible to skip connections. It makes use of batch normalization to boost efficiency while maintaining accuracy.	Implementation is time-consuming.
AlexNet [17]	Unlike a convolutional layer, which depends on local spatial coherence and a narrow receiving field, a fully connected layer learns features from all	Complicated layers with many connections are very computationally costly to create.

	the combinations of the features of the preceding layer.	
VGGNet [11,9]	It only contains 80 percent of the whole number of parameters.	Accuracy decreases in a very progressive manner.
Inception V3 [17]	Allows for any conceivable combination of layers to be used.	Intensive training is required. Cost of calculation is high.

IV. RESULTS AND DISCUSSION

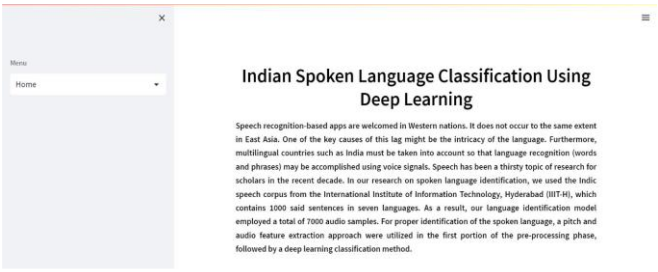

Figure 2. Home Page

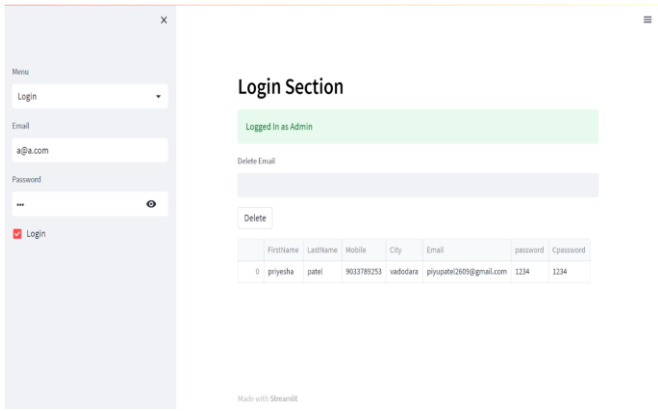

Figure 3. Login Page

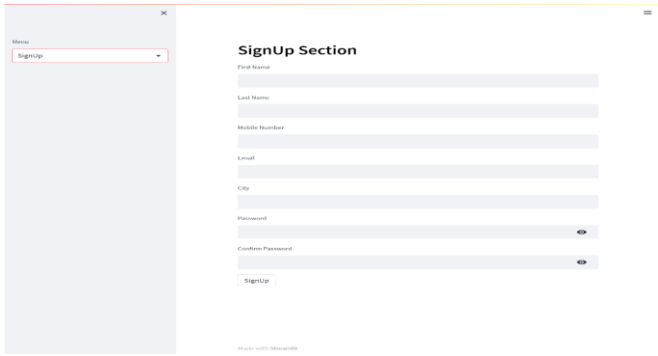

Figure 4. Sign up Page

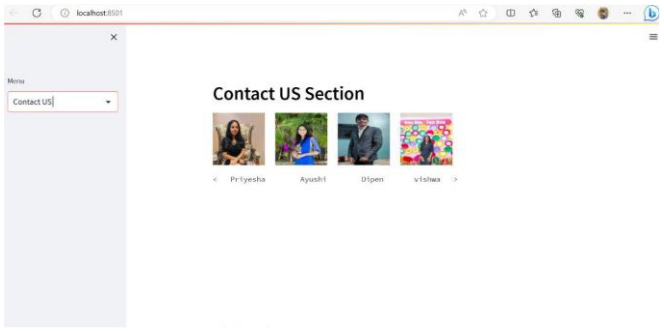

Figure 5. Contact US Page

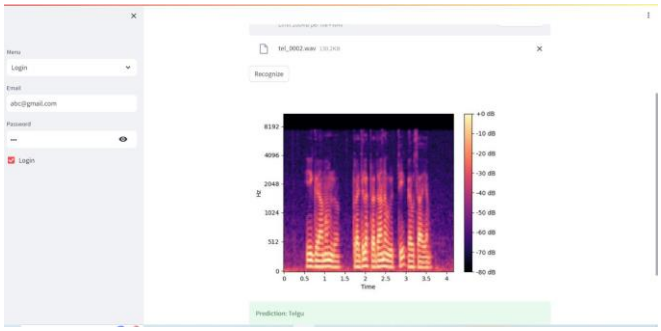

Figure 6. Classification

TABLE II

CLASSIFICATION ANALYSIS

Model	ACC (%)	P (%)	R (%)	F1-score (%)
AlexNet	0.95	0.93	0.93	0.93
VGGNet	0.96	0.95	0.95	0.95
ResNet	0.59	0.52	0.49	0.52
CNN	0.99	0.99	0.99	0.99

V. CONCLUSION

Transfer learning in the realm of audio and voice processing will be facilitated by Spoken Language Recognition in the future. A variety of topologies, as well as unsupervised and supervised training scenarios, should be explored in order to expand the approach further. Using deep feature losses to train deep learning frameworks doesn't have to be confined to just one feature extractor, for example. There may be a wide variety, each customized to capture representations of various components of speech, such as generalized

representations of language, speakers, or distinct linguistic units throughout timelines.

Using Transfer Learning Using CNN Model and Other Different Types of Model in Deep Learning But we got good accuracy in CNN it was 99%.

Using Transfer Learning We Make Audio in Image form Using Spectrogram then we Load dataset and use Transfer Learning To 80 % data go in Training section and 20% Data go for testing and we use AlexNet, VGG, ResNet and CNN Model for Testing Different types of languages and we got 99% Accuracy in CNN Model.

VI.REFERENCES

[1] B. Paul, S. Phadikar, and S. Bera, "Identification Using Deep Learning Approach," pp. 263–274.

[2] H. S. Lee, Y. Tsao, S. K. Jeng, and H. M. Wang, "Subspace-Based Representation and Learning for Phonotactic Spoken Language Recognition," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 28, pp. 3065–3079, 2020, doi: 10.1109/TASLP.2020.3037457.

[3] M. A. A. Albadr and S. Tiun, "Spoken Language Identification Based on Particle Swarm Optimisation–Extreme Learning Machine Approach," Circuits, Syst. Signal Process., vol. 39, no. 9, pp. 4596– 4622, 2020, doi: 10.1007/s00034-020-01388-9.

[4] H. Mukherjee et al., "Deep learning for spoken language identification: Can we visualize speech signal patterns?" Neural Comput. Appl., vol. 31, no. 12, pp. 8483–8501, 2019, doi: 10.1007/s00521-019-04468-3.

[5] S. Gholamdokht Firooz, S. Reza, and Y. Shekofteh, "Spoken language recognition using a new conditional cascade method to combine acoustic and phonetic results," Int. J. Speech Technol., vol. 21, no. 3, pp. 649–657, 2018, doi: 10.1007/s10772-018-9526-5.

[6] D. S. Sisodia, S. Nikhil, G. S. Kiran, and P. Sathvik, "Ensemble learners for identification of spoken languages using mel frequency cepstral coefficients," 2nd Int. Conf. Data, Eng. Appl. IDEA 2020, 2020, doi: 10.1109/IDEA49133.2020.9170720.

[7] G. Singh, S. Sharma, V. Kumar, M. Kaur, M. Baz, and M. Masud, "Spoken Language Identification Using Deep Learning," Comput. Intell. Neurosci., vol. 2021, 2021, doi: 10.1155/2021/5123671.

[8] H. S. Das and P. Roy, A deep dive into deep learning techniques for solving spoken language identification problems. Elsevier Inc., 2019.

[9] N. E. Safitri, A. Zahra, and M. Adriani, "Spoken Language Identification with Phonotactics Methods on Minangkabau, Sundanese, and Javanese Languages," Procedia Comput. Sci., vol. 81, no. May, pp. 182–187, 2016, doi: 10.1016/j.procs.2016.04.047.

[10] P. Heracleous, K. Takai, K. Yasuda, Y. Mohammad, and A. Yoneyama, "Comparative study on spoken language identification based on deep learning," Eur. Signal Process. Conf., vol. 2018- September, pp. 2265–2269, 2018, doi: 10.23919/EUSIPCO.2018.8553347.

[11] R. Fér, P. Matějka, F. Grézl, O. Plchot, K. Veselý, and J. H. Černocký, "Multilingually trained bottleneck features in spoken language recognition," Comput. Speech Lang., vol. 46, pp. 252–267, 2017, doi: 10.1016/j.csl.2017.06.008.

[12] M. Dua, R. K. Aggarwal, and M. Biswas, "Discriminatively trained continuous Hindi speech recognition system using interpolated recurrent neural network language modeling," Neural Comput. Appl., vol. 31, no. 10, pp. 6747–6755, 2019, doi: 10.1007/s00521-018-3499- 9.

[13] O. Giwa and M. H. Davel, "The effect of language identification accuracy on speech recognition accuracy of proper names," 2017 Pattern Recognit. Assoc. South Africa Robot. Mechatronics Int. Conf. PRASA-RobMech 2017, vol. 2018-January,

pp. 187–192, 2017, doi: 10.1109/RoboMech.2017.8261145.

[14] R. W. M. Ng, M. Nicolao, and T. Hain, "Unsupervised crosslingual adaptation of tokenisers for spoken language recognition," Comput. Speech Lang., vol. 46, pp. 327–342, 2017, doi: 10.1016/j.csl.2017.05.002.

[15] M. A. A. Albadr, S. Tiun, M. Ayob, and F. T. AL-Dhief, "Spoken language identification based on optimised genetic algorithm–extreme learning machine approach," Int. J. Speech Technol., vol. 22, no. 3, pp. 711–727, 2019, doi: 10.1007/s10772-019-09621-w.

[16] Y. Ma, R. Xiao, and H. T. B, "An Event-Driven Computational System," vol. 1, pp. 453–461, 2017, doi: 10.1007/978-3-319-70136-3.

[17] P. Beckmann, M. Kegler, H. Saltini, and M. Cernak, "Speech-VGG: A deep feature extractor for speech processing," no. May 2020, 2019, [Online]. Available: http://arxiv.org/abs/1910.09909.

[18] Dhawale, Apurva D., Sonali B. Kulkarni, and Vaishali M. Kumbhakarna. "A Survey of Distinctive Prominence of Automatic Text Summarization Techniques Using Natural Language Processing." In International Conference on Mobile Computing and Sustainable Informatics, pp. 543-549. Springer, Cham, 2020

Cite this article as :

Priyesha Patel, Ayushi Falke, Dipen Waghela, Shah Vishwa, "Utilizing Deep Learning Techniques for the Classification of Spoken Languages in India", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 10, Issue 2, pp.63-69, March-April-2024. Available at doi : https://doi.org/10.32628/CSEIT2390556
Journal URL : https://ijsrcseit.com/CSEIT2390556