A project report on

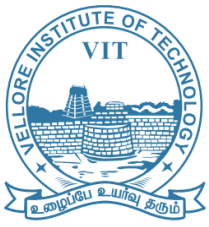
ACCENT RECOGNTION USING DEEP LEARNING TECHNIQUES

Submitted in partial fulfillment for the award of the degree of

B.Tech (Information Technology)

By

ADITYA KOCHERLAKOTA (19BIT0386)



VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF INFORMATION TECHNOLOGY & ENGINEERING

January, 2023

DECLARATION

I here by declare that the thesis entitled “ACCENT RECOGNITION USING DEEP LEARNING TECHNIQUES” submitted by me, for the award of the degree of M.Tech (Software Engineering) is a record of bonafide work carried out by me under the supervision of Prof. Valarmathi B

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore

Date: 26-01-2023 Signature of the Candidate

CERTIFICATE

This is to certify that the thesis entitled "ACCENT RECOGNITION USING DEEP LEARNING TECHNIQUES" submitted by ADITYA KOCHERLAKOTA (19BIT0386), School of Information Technology & Engineering, Vellore Institute of Technology, Vellore for the award of the degree B.Tech (Software Engineering) is a record of bonafide work carried out by him/her under my supervision.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The Project report fulfils the requirements and regulations of VELLORE INSTITUTE OF TECHNOLOGY, VELLORE and in my opinion meets the necessary standards for submission.

Signature of the Guide **Signature of the HoD**

Internal Examiner **External Examiner**

ABSTRACT

Accents refer to a distinctive way of pronouncing a language, especially one associated with a particular country, area, or social class. They typically share characteristics with persons from similar racial, cultural, or social backgrounds. Speech classification is the process of identifying the regional accents of one's fellow people. Speech classification makes it simple to determine a person's ethnicity, personal geography, and religious upbringing. Obviously, if it is done manually, mistakes will occur. Some accents or languages that scarcely anyone speaks or has ever heard of are occasionally mentioned. Since most people don't get it and can't decipher its genuine meaning, this presents a dilemma. This technology can also improve research into speech recognition. To fully utilize the power of convolutional neural network systems previous works have used the concept of a spectrogram image, this allows the audio data to be passed as two-dimensional data and allow the convolutional neural network to improve feature extraction. The purpose of this paper is to compare the performance of other image classification techniques for the purpose of accent recognition whilst using the spectrogram image method.

ACKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to Dr. Valarmathi B, Associate Professor Senior, School of Information Technology & Engineering, Vellore Institute of Technology, for her constant guidance, continual encouragement, understanding; more than all, she taught me patience in my endeavor. My association with her is not confined to academics only, but it is a great opportunity on my part of work with an intellectual and expert in the field of Machine learning.

I would like to express my gratitude to DR.G.VISWANATHAN, Chancellor VELLORE INSTITUTE OF TECHNOLOGY, VELLORE, MR. SANKAR VISWANATHAN, DR. SEKAR VISWANATHAN, MR.G V SELVAM, Vice – Presidents VELLORE INSTITUTE OF TECHNOLOGY, VELLORE, DR. RAMBABU KODALI, Vice – Chancellor, DR. PARTHA SARATHI MALLICK, Pro-Vice Chancellor and Dr. S. Sumathy, Dean, School of Information Technology & Engineering (SITE), for providing with an environment to work in and for his inspiration during the tenure of the course.

In jubilant mood I express ingeniously my whole-hearted thanks to Dr. Usha Devi G., HoD/Professor, all teaching staff and members working as limbs of our university for their not-self-centered enthusiasm coupled with timely encouragements showered on me with zeal, which prompted the acquirement of the requisite knowledge to finalize my course study successfully. I would like to thank my parents for their support.

It is indeed a pleasure to thank my friends who persuaded and encouraged me to take up and complete this task. At last but not least, I express my gratitude and appreciation to all those who have helped me directly or indirectly toward the successful completion of this project.

Place: Vellore

Date: 26-01-2023 Aditya Kocherlaktoa

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

1.2 MOTIVATION

1.3 PROJECT STATEMENT

1.4 OBJECTIVES

1.5 SCOPE OF THE PROJECT

CHAPTER 2

LITERATURE SURVEY

2.1 SUMMARY OF THE EXISTING WORKS

2.2 CHALLENGES PRESENT IN EXISTING SYSTEM

CHAPTER 4

ANALYSIS & DESIGN

4.1 PROPOSED METHODOLOGY

4.2 SYSTEM ARCHITECTURE

4.3 MODULE DESCRIPTIONS

REFERENCES

Chapter 1

INTRODUCTION

1.1 BACKGROUND

Accent classification is identifying and classifying a speaker's accent according to their pronunciation and speech patterns. This is often accomplished by examining the phonetic aspects of a speaker's speech, such as stress patterns, rhythm, intonation, and pronunciation of specific sounds. Accent classification is useful for a wide range of tasks, including speech recognition, language instruction, and linguistics research. Accents can be categorised based on geographical, social, or linguistic characteristics, such as the speaker's original language or region, nation, or ethnic origin.

A speaker's accent is influenced by a variety of circumstances, including their linguistic background, life experiences, and exposure to various languages and accents. Classifying an accent is a complex undertaking that can be difficult. In addition to conventional linguistic analysis techniques, researchers frequently combine them with machine learning algorithms that are trained on vast databases of speech samples to reliably categorise accents.

1.2 MOTIVATION

Accent categorization research can be used to improve speech recognition technology for people with non-standard accents or to better comprehend the linguistic diversity of a population. Accent recognition and detection is a major research subject in speech processing that may be utilized to improve automatic speech recognition.

1.3 PROJECT STATEMENT

The goal of this project is to create a reliable system for classifying accents that reliably classifies a speaker's accent based on speech patterns and pronunciation. The use of machine learning algorithms and a sizable dataset of speech samples from a wide variety of accents will be used to achieve this. Accents will be categorised based on geographic, social, and linguistic aspects, including the speaker's original language and location, nation, and ethnic origin. As a result of this initiative, accents will be represented in various fields more accurately and in a wider variety, which will benefit speech recognition, language acquisition, and linguistics research.

1.4 OBJECTIVES

The objectives of this project are two-fold. The first is to use spectrogram images as input for the algorithms. The second is to compare the results of the different algorithms employed. In this project, the algoritms used will be the basic MLP (Multi Layered Perceptrons), the popular CNN (Convolutional Neural Networks), and the relatively new ViT (Vision Transformers). All the algorithms are popular deep learning methods for Image classification, through this comparison we can find the best perfomer for the usecase of Accent Classification.

1.5 SCOPE OF THE PROJECT

The scope of the accent classification project includes the following:

Data collection: Gathering a large and diverse dataset of speech samples from a range of accents.

Feature extraction: Extracting relevant phonetic features from the speech samples, such as stress patterns, rhythm, intonation, and pronunciation of individual sounds.

Algorithm development: Developing machine learning algorithms to analyze the extracted features and accurately classify the accent of a speaker.

Model evaluation: Evaluating the performance of the accent classification system through testing and analysis.

Documentation: Documenting the process and methodology used in the project, as well as the results and any limitations of the accent classification system.

This project will focus on developing an effective and accurate accent classification systems, but will not include developing or improving speech recognition or language learning applications themselves. The accent classification system developed in this project will be a standalone tool

Chapter 2

LITERATURE SURVEY

2.1 SUMMARY OF THE EXISTING WORKS

SNO	Title	Merits	Demerits
1	Accent Recognition Using a Spectrogram Image Feature-Based Convolutional Neural Network	takes advantage of the convolutional neural networks' ability to characterize two-dimensional signals	Transfer learning was used
2	Automatic accent classification of foreign accented Australian English speech	Does not require manually labelled data	Low number of classes
3	An empirical study of automatic accent classification	Large number of classes	Low detection rate
4	Foreign English Accent Classification Using Deep Belief Networks	Much better accuracy than SVM,K-NN and random forest	Requires large amounts of data for good results
5	Accent classification using Machine learning and Deep Learning Models	Good comparison between ML and DL techniques	CNN was not effectively used.
6	Speaker Accent Classification System using Fuzzy Canonical Correlation-Based Gaussian Classifier	Novel method with focus on minimizing the distance between the cluster centroids, but also maximizing the out-of-class variations.	Comparatively a slower algorithm because we have to compute the membership of each data point in each cluster.
7	Australian Accent-Based Speaker Classification,	Novel method that used fusion of classifiers, which showed high	

		performance	
8	Accent classification in speech	Novel approach that tested on the formant frequenceies of the accent markers	Low Number of classes
9	Features of Speech Audio for Accent Recognition	Used MFCCs for feature extraction which improved Accuracy	Low number of layers in CNN
10	Dialect/accent classification via boosted word modeling,	Uses a novel approach of converting the problem to a word-based dialect classification problem	Low number of classes
11	VFNet: A Convolutional Architecture for Accent Classification	they presented VFNet (Variable Filter Net), a convolutional neural network (CNN) based architecture which captures a hierarchy of features to beat the previous benchmarks of accent classification, hrough a novel and elegant technique of applying variable filter sizes along the frequency band of the audio utterance	
12	Speaker Accent Recognition Using Machine Learning Algorithms	Used many types of algorithms and created a good comparison	
13	Text-Independent Foreign Accent Classification using Statistical Methods	They investigate statistical approaches	

		which differ from the a priori knowledge they need: GMM, which requires neither phonetic knowledge nor labelling, phone recognition (without a lexicon), sentence recognition (with a lexicon and a grammar).	
14	Improved accent classification combining phonetic vowels with acoustic features	they combined phonetic knowledge, such as vowels, with enhanced acoustic features to build an improved accent classification system	Demanding pre processing steps
15	A Machine Learning Approach to Recognize Speakers Region of the United Kingdom from Continuous Speech Based on Accent Classification	High accuracy due to the MFCC feature extraction, Comparative analysis of three ML algorithms	
16	MARS: A Hybrid Deep CNN-based Multi-Accent Recognition System for English Language,	Used and created their own database	Only 6 states
17	US Accent Recognition Using Machine Learning Methods	Comparative analysis on the use of SVM and KNN	
18	An evolutionary approach for accent classification in IVR systems	Built to handle imbalanced data	Low accuracy
19	Speaker accent recognition through statistical descriptors of Mel-bands	Novel approach with High accuracy	Low number of classes

	spectral energy and neural network model		
20	Accent Classification Using Support Vector Machines,	High accuracy	Low number of classes, connot handle imbalanced data well

2.2 CHALLENGES PRESENT IN EXISTING SYSTEM

Feature extraction is a demanding and time-consuming process. Using the spectrogram method in conjunction with various image classification algorithm, the process of feature extraction is simpli fied. The proposed system focuses on regional accent recognition, a more challenging problem than non-native accent recognition. The process of using spectrogram images also allows inputs to be two-dimensional in nature, which might prove to provide better results.

Chapter 4

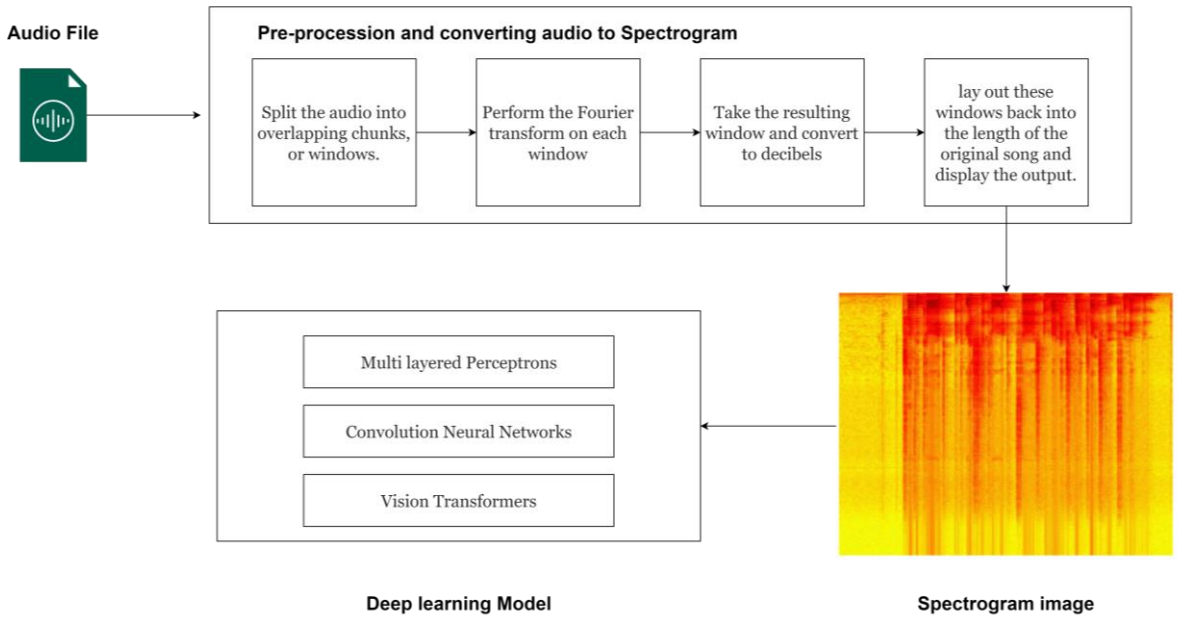
ANALYSIS & DESIGN

4.1 PROPOSED METHODOLOGY

This paper proposes the use of deep learning techniques to classify the different accents in a region. The Deep learning techniques will be taking spectrogram images as inputs. This will help simplify the process of feature ectractions, which otherwise considered demanding and time-consuming [1].

The deep learning techniques used will be namely MLPs, CNNs and ViT, their performance will compared to find the best performing algorithm for this use case.

4.2 SYSTEM ARCHITECTURE



4.3 MODULE DESCRIPTIONS

4.3.1 Spectrogram Image

A spectrogram is a graphical representation of the frequency spectrum of a signal over time. It is a 2-dimensional image that displays the frequency content of a signal as it evolves over time, making it a useful tool for analyzing and understanding various signals and signals processing applications.

The x-axis of a spectrogram represents time, the y-axis represents frequency, and the amplitude of the signal at each time-frequency point is represented by the color or grayscale intensity. The color or grayscale intensity typically indicates the strength or power of the signal at that particular time-frequency point, with brighter or higher intensity values representing stronger signals and darker or lower intensity values representing weaker signals.

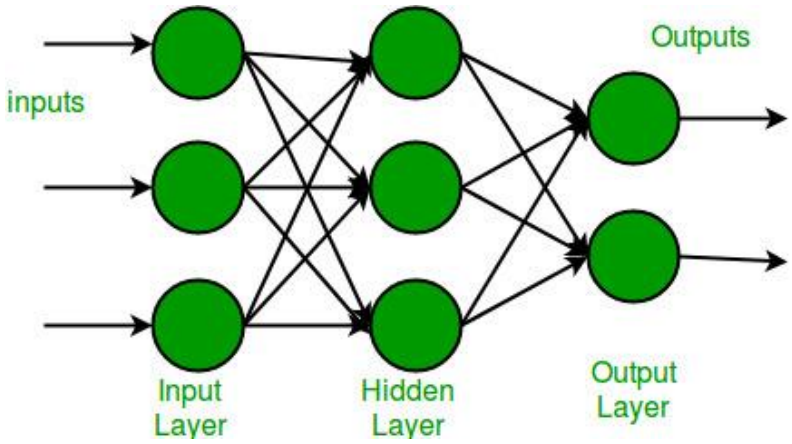
One of the main uses of spectrograms is in the analysis of speech signals. By examining the spectrogram of a speech signal, linguists and speech scientists can identify the frequency components that make up speech sounds, such as vowels and consonants, and gain insights into the production and perception of speech.

To create a spectrogram, a signal is typically first transformed into the frequency domain using techniques such as the Fourier transform or the Short-Time Fourier Transform (STFT). The frequency domain representation of the signal is then divided into overlapping segments, and the magnitude of the signal in each segment is computed and displayed as a function of time and frequency.

4.3.2 Multi-Layered Perceptrons (MLPs)

Multi-Layer Perceptrons (MLPs) are a type of artificial neural network that is commonly used for supervised learning tasks such as classification and regression. They are called Multi-Layer Perceptrons because they consist of multiple layers of artificial neurons, or "perceptrons". Each layer of the MLP consists of a set of artificial neurons, and each neuron receives input from the neurons in the previous layer, processes the input using an activation function, and outputs a signal to the neurons in the next layer. The final layer of the MLP outputs the final prediction or classification of the input data. MLPs are trained using a process called backpropagation, where the error between the predicted output and the actual output is calculated, and the weights of the

neurons are adjusted to minimize the error. This process is repeated many times until the MLP is able to make accurate predictions on the training data. One of the key advantages of MLPs is their ability to learn non-linear relationships between inputs and outputs, as opposed to linear models such as linear regression. This makes them well-suited for complex data and tasks, such as image classification and speech recognition. However, MLPs can be prone to overfitting, where the network becomes too complex and performs well on the training data but poorly on unseen data. To combat this, techniques such as regularization and early stopping can be used.
To perform image classification using MLP, the input image is typically pre-processed to extract relevant features, such as edges and textures. These features are then fed into the MLP as input, and the network uses its multiple layers of neurons to process the features and make a prediction.
The MLP is trained on a labeled dataset of images, where the goal is to learn the relationships between the features of the images and their corresponding class labels. The training process involves using an optimization algorithm, such as gradient descent, to minimize the error between the predicted class labels and the actual class labels.



Architecture of an MLP

4.3.3 Convolutional Neural Networks (CNNs)

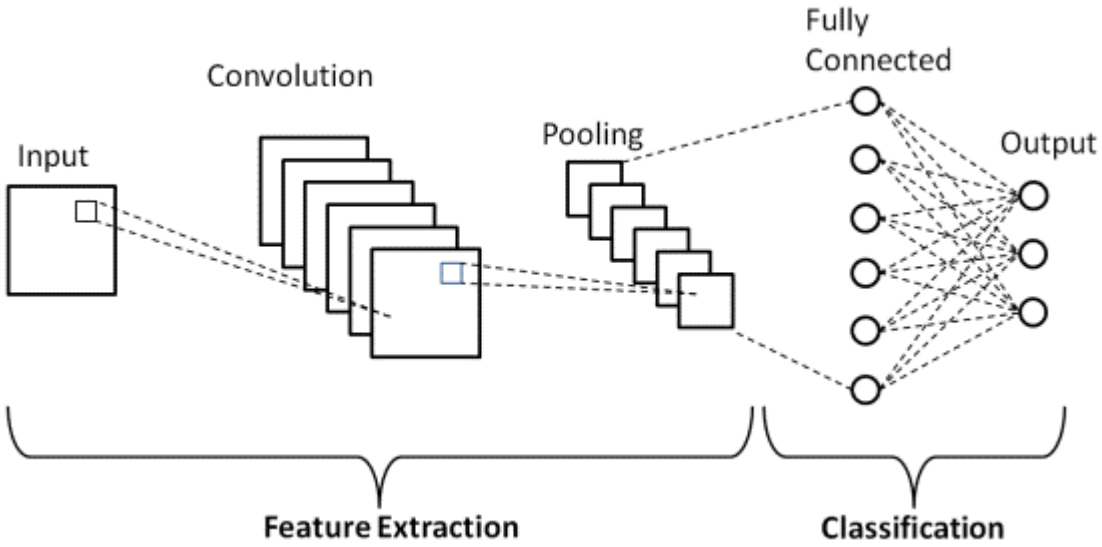
Convolutional Neural Networks (CNNs) are a type of neural network designed specifically for image classification tasks. They are called Convolutional Neural Networks because they use a technique called convolution to extract meaningful features from the input image. In image classification, the goal is to predict the class label of an input image. With CNNs, the input image is processed through multiple layers of neurons, each of which performs a convolution operation. The convolution operation involves applying a small filter to each portion of the image, producing a feature map that captures the important patterns and structures in the image. The output of the convolution operation is then passed through activation functions, which introduce non-linearity into the network. The resulting feature maps are then processed through multiple fully connected layers, which learn the relationships between the features and the class labels.

One of the key advantages of CNNs is their ability to learn hierarchical representations of the input data. This allows the network to automatically learn and extract increasingly complex

features from the input image, such as edges, textures, and shapes, and use them to make accurate predictions.

Another advantage of CNNs is their ability to handle large amounts of data, as they can be trained on large datasets of images using GPUs, which allows for faster processing.

In conclusion, Convolutional Neural Networks are a powerful and widely used approach for image classification tasks. By using convolution to extract meaningful features from the input image, and learning the relationships between the features and the class labels, CNNs can make accurate predictions and handle large amounts of data.



Architecture of a CNN

4.3.4 Vision Transformers (ViT)

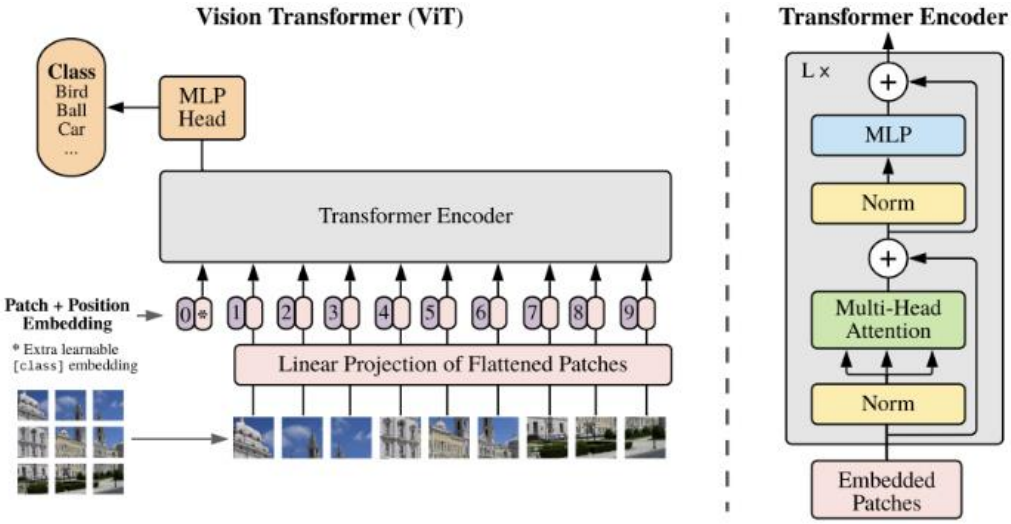
Vision Transformers are a type of deep learning model that are based on the Transformer architecture, originally developed for natural language processing tasks. They have been adapted for computer vision tasks, such as image classification and object detection, by incorporating the key elements of the Transformer architecture into the neural network structure.

In a Vision Transformer, the input image is divided into a set of non-overlapping patches, which are then flattened and fed into the Transformer as input tokens. The Transformer processes the tokens using self-attention mechanisms, which allow the network to focus on different regions of the image and learn the relationships between the regions.

One of the key advantages of Vision Transformers is their ability to process images in a fully parallel manner, without the need for sequential processing like in traditional Convolutional

Neural Networks (CNNs). This allows Vision Transformers to handle larger images, and to make predictions faster.

Another advantage of Vision Transformers is their ability to learn global representations of the input image, allowing the network to capture the overall context and relationships between different regions of the image.



Architecture of Vision Transformers

REFERENCES

1. Cetin, O. Accent Recognition Using a Spectrogram Image Feature-Based Convolutional Neural Network. Arab J Sci Eng (2022). https://doi.org/10.1007/s13369-022-07086-9 I nternational Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

2. K. Kumpf and R. W. King, "Automatic accent classification of foreign accented Australian English speech," Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, Philadelphia, PA, USA, 1996, pp. 1740-1743 vol.3, doi: 10.1109/ICSLP.1996.607964., 1 (March 2018), 93–120. https://doi.org/10.1007/s10772-018-9491-z

3. G. Choueiter, G. Zweig and P. Nguyen, "An empirical study of automatic accent classification," 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 2008, pp. 4265-4268, doi: 10.1109/ICASSP.2008.4518597.

4. R. Upadhyay and S. Lui, "Foreign English Accent Classification Using Deep Belief Networks," 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2018, pp. 290-293, doi: 10.1109/ICSC.2018.00053..

5. A. Purwar, H. Sharma, Y. Sharma, H. Gupta and A. Kaur, "Accent classification using Machine learning and Deep Learning Models," 2022 1st International Conference on Informatics (ICI), Noida, India, 2022, pp. 13-18, doi: 10.1109/ICI53355.2022.9786885.NY, USA, 801–804. https://doi.org/10.1145/2647868.2654984

6. S. Ullah and F. Karray, "Speaker Accent Classification System using Fuzzy Canonical Correlation-Based Gaussian Classifier," 2007 IEEE International Conference on Signal Processing and Communications, Dubai, United Arab Emirates, 2007, pp. 792-795, doi: 10.1109/ICSPC.2007.4728438.

7. P. Nguyen, D. Tran, X. Huang and D. Sharma, "Australian Accent-Based Speaker Classification," 2010 Third International Conference on Knowledge Discovery and Data Mining, Phuket, Thailand, 2010, pp. 416-419, doi: 10.1109/WKDD.2010.80.

8. S. Deshpande, S. Chikkerur and V. Govindaraju, "Accent classification in speech," Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05), Buffalo, NY, USA, 2005, pp. 139-143, doi: 10.1109/AUTOID.2005.10

9. Y. Singh, A. Pillay and E. Jembere, "Features of Speech Audio for Accent Recognition," 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 2020, pp. 1-6, doi: 10.1109/icABCD49160.2020.9183893.

10. Rongqing Huang and J. H. L. Hansen, "Dialect/accent classification via boosted word modeling," Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., Philadelphia, PA, USA, 2005, pp. I/585-I/588 Vol. 1, doi: 10.1109/ICASSP.2005.1415181.

11. A. Ahmed, P. Tangri, A. Panda, D. Ramani and S. Karmakar, "VFNet: A Convolutional Architecture for Accent Classification," 2019 IEEE 16th India Council International Conference (INDICON), Rajkot, India, 2019, pp. 1-4, doi: 10.1109/INDICON47234.2019.9030363.

12. A. A. Ayrancı, S. Atay and T. Yıldırım, "Speaker Accent Recognition Using Machine Learning Algorithms," 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, 2020, pp. 1-6, doi: 10.1109/ASYU50717.2020.9259902.

13. D. Fohr and I. Illina, "Text-Independent Foreign Accent Classification using Statistical Methods," 2007 IEEE International Conference on Signal Processing and Communications, Dubai, United Arab Emirates, 2007, pp. 812-815, doi: 10.1109/ICSPC.2007.4728443.

14. Z. Ge, "Improved accent classification combining phonetic vowels with acoustic features," 2015 8th International Congress on Image and Signal Processing (CISP), Shenyang, China, 2015, pp. 1204-1209, doi: 10.1109/CISP.2015.7408064.

15. M. F. Hossain, M. M. Hasan, H. Ali, M. R. K. R. Sarker and M. T. Hassan, "A Machine Learning Approach to Recognize Speakers Region of the United Kingdom from Continuous Speech Based on Accent Classification," 2020 11th International Conference on Electrical and Computer Engineering (ICECE), Dhaka, Bangladesh, 2020, pp. 210-213, doi: 10.1109/ICECE51571.2020.9393038.

16. S. Darshana, H. Theivaprakasham, G. Jyothish Lal, B. Premjith, V. Sowmya and K. Soman, "MARS: A Hybrid Deep CNN-based Multi-Accent Recognition System for English Language," 2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR), Hyderabad, India, 2022, pp. 1-6, doi: 10.1109/ICAITPR51569.2022.9844177.

17. M. Muttaqi, A. Degirmenci and O. Karal, "US Accent Recognition Using Machine Learning Methods," 2022 Innovations in Intelligent Systems and Applications Conference (ASYU), Antalya, Turkey, 2022, pp. 1-6, doi: 10.1109/ASYU56188.2022.9925265.

18. S. Ullah and F. Karray, "An evolutionary approach for accent classification in IVR systems," 2008 IEEE International Conference on Systems, Man and Cybernetics, Singapore, 2008, pp. 418-423, doi: 10.1109/ICSMC.2008.4811311.

19. Y. Ma, M. Paulraj, S. Yaacob, A. Shahriman and S. K. Nataraj, "Speaker accent recognition through statistical descriptors of Mel-bands spectral energy and neural network model," 2012 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT), Kuala Lumpur, Malaysia, 2012, pp. 262-267, doi: 10.1109/STUDENT.2012.6408416.

20. C. Pedersen and J. Diederich, "Accent Classification Using Support Vector Machines," 6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007), Melbourne, VIC, Australia, 2007, pp. 444-449, doi: 10.1109/ICIS.2007.47