# Investigation of Advancements in Machine Learning Algorithms for Accented Speech Recognition

Yuhao Guo
*International Public Policy, University College London, London, U.K.*

Keywords: Machine Learning, Review.

Abstract: The rapid advancement in technology has catalyzed the development of machine learning models, significantly impacting various domains, including speech recognition systems. This paper delves into the historical progression and transformation of machine learning models, emphasising on the application of machine learning algorithms within the domain of Voice Activity Detection (VAD) speech recognition systems. The narrative transitions from a broad overview of machine learning frameworks to a focused discussion on speech recognition technologies, specifically addressing the challenge of accentual language recognition. Through the lens of Mel-Frequency Cepstral Coefficients (MFCC) application, the paper dissects the learning models and elucidates the segmentation and recognition processes integral to speech recognition systems. Various algorithms are briefly reviewed to underscore the ongoing enhancements and the scholarly contributions towards refining these models. However, the paper also acknowledges the inherent limitations in the current understanding and application of these models. It points out the superficial treatment of certain mainstream machine learning models and their algorithms, alongside the potential obsolescence due to rapid technological evolution. The exposition concludes by recognizing the potential gaps in the depth of understanding specific to VAD and MFCC processes, attributed to the author's academic limitations, which may impinge on the thoroughness of algorithmic familiarity.



## 1 INTRODUCTION

Accents are distinctive speech patterns associated with speakers from various linguistic backgrounds. Speakers with accents usually share standard features in their speech patterns with people from similar ethnic, cultural, or social backgrounds (Purwar et al. 2022). English is widely spoken throughout the world, by a quarter of the world's population. The challenge of phonetic categorization involves identifying the various accent types that compatriots exhibit across different geographical locations (Salomone & Salomone 2022). The rapid development of globalization has intensified communication barriers across different parts of the world, presenting a challenge that needs to be addressed. Language accents, including those in English, significantly impact education and particularly affect the language education and training of young people, leading to misunderstandings and disconnections.

Many studies have researched and made substantial progress in improving speech recognition accuracy, especially accent recognition. For instance, Siddhant et al. have attempted to develop systems that can recognise speech accents more accurately. The results demonstrated that the decision tree performs best among various machine learning models, even when compared to deep learning models such as Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM), with an accuracy of 97.36% (Purwar et al. 2022). They have favoured careful accent categorisation and identification, with experimental simulations constantly using different accent data; it also suggests solutions to specific problems, particularly the loss of detail in decoder outputs (Purwar et al. 2022). Singh et al. have focused on automatic speech recognition and believe that its application in education is excellent; they have designed a new speech recognition system, developed and tested a deep concatenated network and found a 15.4% performance improvement over existing techniques. At the same time, they carried out a detailed analysis of speech recognition errors (Singh et al. 2020). In addition to all the above work on the systematic study of linguistic accents, Shergill et al.



Guo, Y.
Investigation of Advancements in Machine Learning Algorithms for Accented Speech Recognition.
DOI: 10.5220/0012866900004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 201-205
ISBN: 978-989-758-690-3

have developed specific methods for setting up and evaluating accent analysis. Accents vary in terms of sound quality, phoneme articulation and prosody. Because it is challenging to extract these precise features, they concluded that existing work uses alternative features, such as spectral features that capture the frequency of speech, including five significant features: Mel Falling Spectral Coefficient (MFCC), spectrogram, chromaticity map, spectral centroid and spectral roll-off (Shergill et al. 2020). They have researched and proposed a new way of capturing accents to improve the recognition efficiency of the system. These features can improve the accuracy of accent categorisation for accented language systems. Other researchers have categorised user input capture in the form of speech data. They are modelling human speech accents and gender recognition as classification tasks. They used a deep convolutional neural network and experimentally developed an architecture that maximises classification accuracy for the above functions. Gender categorisation was more straightforward to predict with high accuracy than accent. The categorisation of accents was complex because the overlap of regional accents prevented them from being categorised with high accuracy (Najafian et al. 2016). Deng et al. have carefully delineated the field of automatic speech recognition within a framework such as improved accent recognition and accented speech recognition within a self-supervised learning framework (Deng et al. 2021). Some other studies have categorised accents, particularly the origin of English accents, to classify them for better identification (Ai et al 2008).

This paper presents a comprehensive review of the application of artificial intelligence models in accented speech recognition systems. Section 2 outlines the workflow of various relevant methodologies. A discussion on the findings and implications of these applications is provided in Section 3. The paper concludes in Section 4 with a summary of the key insights and contributions of the research.

# 2 METHOD

## 2.1 Framework of Developing Machine Learning Model

A prevailing notion within the philosophy of science posits that models which are simplified and idealized are more comprehensible than their complex or abstract counterparts (Bokulich 2008). The landscape of artificial intelligence (AI) architecture is continuously evolving, leading to the development of numerous machine learning models, with simpler models often being more straightforward to grasp. These models are particularly adept at addressing "what-if" scenarios or exploring causal relationships, thereby effectively identifying and underscoring critical distinctions. Deep Neural Networks (DNNs) serve as a prime example of this. They fundamentally operate by leveraging vast datasets to perform classifications, predictions, and inferences. Specifically, DNNs process extensive data inputs to produce representations that facilitate generalized predictions, enabling the anticipation of future events (Sullivan 2022). The most essential steps in machine modelling are model building, data collection and model management. Building machine learning models is an iterative process. Data collection is becoming one of the critical bottlenecks. It is well known that running machine learning end-to-end, including collection, cleaning, analysis, visualisation and feature engineering. But one problem that arises with the emergence of new machine learning models is the lack of data, especially the lack of training data such as Deep learning, which may require a lot of training data to perform well. The simple method of manual labelling can be used when there is a lack of training data, but it is expensive and requires domain expertise (Roh et al. 2021). There are three main approaches to data collection. First, if the goal is to share and search for new datasets, data collection techniques can be used to discover, extend, or generate datasets. Second, once a dataset is available, various data labelling techniques can be used to annotate individual examples. Finally, rather than labelling new datasets, existing data can be improved or trained on well-trained models. These three approaches are not necessarily distinct and can be used simultaneously (Roh et al. 2021).

Management of the machine learning model is equally important. It is impossible to manage models that have moved on over time sustainably. At the same time, the management of the model is equally important. For example, Manasi Vartak et al. contributed to model management by developing a new software management model database for management. To automate the tracking of machine learning models, the back-end introduces a standard abstraction layer to represent models and pipelines. The ModelDB front-end allows for visual exploration and analysis of models through a web-based platform. The management of machine learning models is achieved by visual exploration and analysis of models through a web-based interface (Vartak et al. 2016).

## 2.2 Algorithms for Accented Speech Recognition

Voice modelling, on the other hand, operates in a slightly different way than traditional machine learning. The voice activity detectors (VADs) based on statistical models have shown impressive performances, mainly when fairly precise statistical models are employed. However, the collection of speech data is affected by the environment. In particular, technical barriers exist to working in extreme noise condition (Ramirez et al. 2004). Therefore, further improvements in the algorithm are needed. A new VAD algorithm is proposed by Javier Ramírez et al. To improve the robustness of voice detection and the performance of speech recognition systems in noisy environments. The algorithm measures the long-term spectral dispersion (LTSD) between speech and noise. It formulates speech/non-speech decision rules by comparing the long-term spectral envelope with the average noise spectrum, resulting in highly discriminative decision rules and minimising the average number of decision errors (Ramirez et al. 2004). Jong Won Shin et al. propose more complex and accurate computational models. They used the generalised gamma distribution (GΓD) as a new model for VAD based on the likelihood ratio test. A parameter estimation algorithm is proposed based on the maximum likelihood principle (Shin et al. 2010). A VAD algorithm is proposed, where the LRT is modelled based on parameters represented by a generalised gamma distribution (GΓD).

The above is a process of performing speech data collection, machine learning model building, and algorithmic rules. It is mainly used in the medical field, focusing on analysing sounds made in the body. Based on the sound analysis is further subdivided into the collection of sound data, i.e., building and analysing the learning model of the human voice. The use of MFCC features shown in Fig. 1 for the speaker accent recognition system is based on the analysis of patterns such as the speaker's manner of speaking and the choice of words used while speaking, the information about the parameters contained in the human voice and recognition of them at the highest possible rate and the algorithm is updated. Ahmet Aytuğ AYRANCI et al. used 9 ML classification algorithms through the MFCC speaker accent recognition system. In addition, a k-fold cross-validation technique was used to test the dataset independently. In this way, the performance of ML algorithms is demonstrated when the dataset is divided into k parts (AYRANCI et al. 2021).

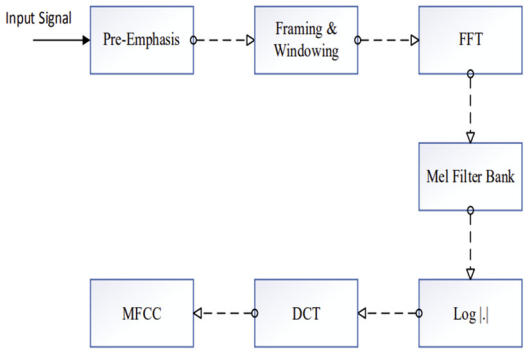


Figure 1: The diagram of MFCC technique (AYRANCI et al. 2021).

For data, different accents are automatically categorised into different datasets through the MFCC system, and algorithms form the result (AYRANCI et al. 2021). In addition, Multi-Layer Perceptron (MLP) algorithm shown in Fig. 2 is used in classification and regression problems.

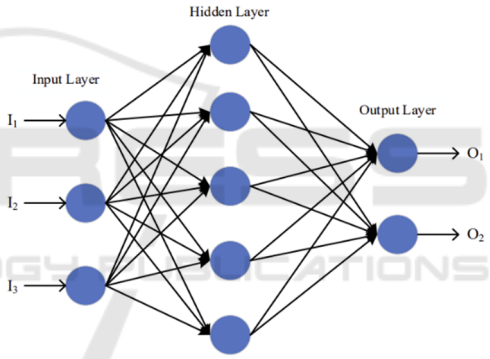


Figure 2: The architecture of the MLP (AYRANCI et al. 2021).

# 3 DISCUSSION

However, there is still a lot of room for improvement regarding the machine learning model-based speech recognition systems that exist today, and it is also evident that there is a continuous improvement in the algorithms. However, operational logic is generally very complex, and the learning models for speech still present a significant challenge regarding training data. Faced with the problem of how to deal with the training data, Javier Ramírez et al.are focusing on the algorithmic problem of learning the model, i.e., how to carry out an accurate analysis of the collected data, proposing new algorithms to process the data to improve the accuracy of the data output. Jong Won

Shin et al. then upgraded the collection side of speech data and innovated the original VAD model.

Based on the above findings, the voice is further classified, i.e., the human accent, which is recognised with the help of MFCC systems and algorithms (e.g., MLP, Random Forest (RF), Decision Tree (DT), Radial Basis Function (RBF), k-Nearest Neighbour (k-NN), Plain Bayes (NB) and Logic Model Tree (LMT)). That is, automatic speech recognition (ASR) systems. And accent recognition can improve the performance of many automatic speech recognition (ASR) systems (Najafian et al. 2016). However, there is some debate about how to perform accent recognition. Ahmet Aytuğ AYRANCI then tends to algorithmic enhancement of the MFCC system, and choosing the correct algorithm will increase the efficiency of MFCC. Justina Grigaliūnaitė, on the other hand, prefers to use 2D spectrogram representation and ResNet model for sound data classification rather than the widely used MFCC (Grigaliūnaitė 2022). The MFCC is more widespread than the VDS system developed for the medical field. It also has better future prospects, as it can be used both for differentiating between different countries (English pronunciation can vary from country to country) and can be helpful to customs borders. It can also help people from different countries learn a language or improve their learning efficiency. Machine learning algorithms, particularly Support Vector Machines (SVMs) and Random Forests can play a crucial role in accent classification when applied to appropriate training sets (Thakkar et al. 2019).

Particularly about e-learning, the MFCC system is used to analyse the transformation to generate accent datasets and, thus, better subtitle generation. For example, designing software to generate software to recognise the teacher's accent and thus generate subtitles to help students learn as shown in Fig. 3. It is crucial first to identify a machine learning model with input and output types.

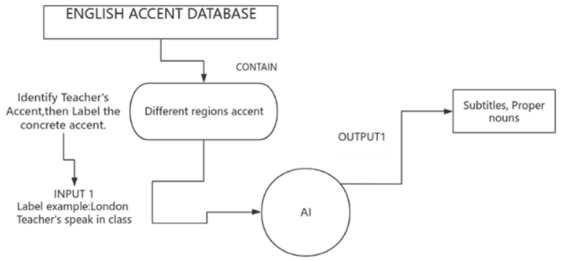


Figure 3: The potential system for the teacher's accent speech recognition (Picture credit : Original).

In terms of the data generation, it is derived by modelling the accent data already processed by the algorithm as shown in Fig. 4. The data is then categorised to generate subsets by the MLP algorithm or a combination of several algorithms.

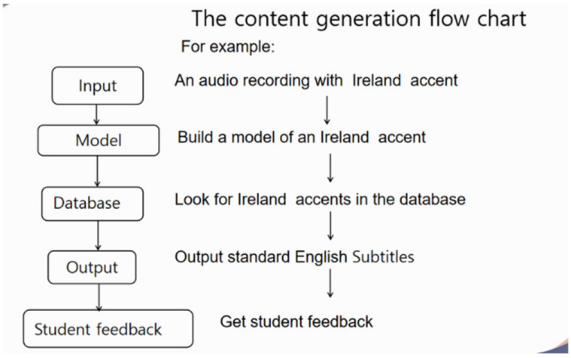


Figure 4: The content generation flow chart (Picture credit: Original).

## 4 CONCLUSION

This paper provides an overview of the evolution of machine learning models from primitive algorithms to DNNs and the broader application of DNNs in fields such as the VAD speech recognition system. It transitions from general machine learning frameworks to specific speech recognition systems, summarizing the learning model for accentual language recognition through MFCC application and detailing the recognition process within the speech recognition system's segmentation framework. The document briefly outlines various algorithms, highlighting the improvements and enhancements contributed by different scholars. Nevertheless, it acknowledges limitations, including a cursory introduction to a few mainstream machine learning models, their algorithms, and the fundamental processes. Constraints related to time may render some discussed algorithms or models obsolete. Additionally, the depth of understanding of specific processes related to VAD and MFCC might be limited due to the author's academic background, potentially affecting familiarity with each algorithm.

## REFERENCES

A. Purwar, H. Sharma, Y. Sharma, H. Gupta, A. Kaur, *Accent classification using Machine learning and Deep Learning Models*, in Proc. 2022 1st Int. Conf. on

Informatics (ICI), Noida, India, pp. 13-18, doi: 10.1109/ICI53355.2022.9786885 (2022)

R. Salomone, R.C. Salomone, Oxf. Univ. Press (2022)

Y. Singh, A. Pillay, E. Jembere, *Features of Speech Audio for Accent Recognition*, in Proc. Int. Conf. on Artif. Intell., Big Data, Comput. and Data Commun. Syst. (icABCD), Durban, South Africa, pp. 1-6, doi: 10.1109/icABCD49160.2020.9183893 (2020).

J. Shergill et al. *Accent and gender recognition from English language speech and audio using signal processing and deep learning.* InHybrid Intelligent Systems: 20th International Conference on Hybrid Intelligent Systems (HIS 2020), December 14-16, 2020 2021 (pp. 62-72).

S. Najafian, S. Safavi, J.H.L. Hansen, M. Russell, *Improving speech recognition using limited accent diverse British English training data with deep neural networks*, in Proc. IEEE Int. Workshop on Mach. Learn. for Signal Process. (MLSP), Vietri sul Mare, Italy, pp. 1-6, doi: 10.1109/MLSP.2016.7738854 (2016)

Deng K, Cao S, Ma L. Improving accent identification and accented speech recognition under a framework of self-supervised learning. arXiv preprint arXiv:2109.07349. (2021) https://arxiv.org/abs/2109.07349

L. Ai, S.-Y. Jeng, H. Beigi, A New Approach to Accent Recognition and Conversion for Mandarin Chinese, arXiv:2008.03359 [eess.AS], https://doi.org/10.48550/arXiv.2008.03359.

A. Bokulich, Camb. Univ. Press (2008)

E. Sullivan, The Brit. J. for the Philos. of Sci., **73**(1), March 2022, https://doi.org/10.1093/bjps/axz035

Y. Roh, G. Heo, S.E. Whang, IEEE Trans. on Knowl. and Data Eng., **33**(4), pp. 1328-1347, doi: 10.1109/TKDE.2019.2946162, 1 April 2021.

M. Vartak, H. Subramanyam, W.-E. Lee, S. Viswanathan, S. Husnoo, S. Madden, M. Zaharia, MODELDB: A System for Machine Learning Model Management. MIT, https://people.eecs.berkeley.edu/~matei/papers/2016/hilda_modeldb.pdf (2016)

J. Ramírez, et al. Speech Commun., 42(3–4), 271-287, https://doi.org/10.1016/j.specom.2003.10.002 (2004).

J.W. Shin, J.H. Chang, N.S. Kim, Comput. Speech Lang., **24**(3), 515-530, doi: 10.1016/j.csl.2009.02.003 (2010)

A.A. AYRANCI, S. ATAY, T. YILDIRIM, Int. J. Adv. Eng. Pure Sci., **33**, 17-27,

J. Grigaliūnaitė, Vilniaus universitetas (2022)

M. Thakkar, S. Elias, A. Ashok, *Speech Recognition Learning Framework for Non-Native English Accent*, in Proc. Int. Conf. on Data Sci. and Eng. (ICDSE), Patna, India, pp. 84-89, doi: 10.1109/ICDSE47409.2019.8971486 (2019).