Ibn Al-Haitham Journal for Pure and Applied Sciences
Journal homepage: jih.uobaghdad.edu.iq
PISSN: 1609-4042, EISSN: 2521-3407
IHJPAS. 2026; 39(2)

English Accent Classification Using Deep Learning Techniques

Sarah Jassim Ahmed[1*], and Husam Ali Abdulmohsin[2]
[1,2]Computer Science Department, College of Science, University of Baghdad, Baghdad, Iraq.
*Corresponding Author.

Received: 10/ August/2025
Accepted:21/December/2025
Published: 20/April/2026
doi.org/10.30526/39.2.4277

Abstract

In recent years, significant advancements have been made in deep learning technology within the field of speech applications, which has resulted in an increased interest in accent classification. The growing need for accurate speech recognition technology requires enhancing the ability of machines to identify accents, which results in giving a critical challenge in speech processing. The variety of English accents poses significant difficulties for automated speech recognition (ASR) systems, adversely impacting transcription accuracy and speaker intelligibility. This study aims to address this challenge by developing a deep learning model efficient in accurately classifying regional English dialects throughout the United Kingdom and Ireland. The proposed system combines a one-dimensional Convolutional Neural Network with a Gated Recurrent Unit (1D-CNN-GRU) architecture and utilizes Mel-Frequency Cepstral Coefficients (MFCCs) as acoustic features. The UK and Ireland English Dialect (UIED) dataset, consisting of 17,877 recordings across six accent categories (Welsh, Northern, Southern, Scottish, Irish, and Midlands English), was utilized for assessment. Experimental results indicate that the proposed model surpasses previous techniques, with an accuracy of 98.71%, hence underscoring its efficacy in capturing accent-specific temporal and spectral patterns. The findings improve the development of accent-resistant ASR systems and establish a basis for future research using transformer-based embedding and prosodic characteristics.

Keywords: Accent classification, Deep learning, English language, speech recognition.

1. Introduction

The accent is a considerable challenge in communication for most spoken languages. "Accent" refers to the diversity in the pronunciation of words among speakers of the same language[1]. It is a speech pattern seen in a speaker's language. The most prominent feature of non-native speech is accentedness. This feature derives from variations between the features of native and non-native languages[2]. An accent is described phonetically as a distinct pronunciation formed by the phonetic features of the speaker's native language, which are transferred into their second language speech [3]. Moreover, depending on the fluency of the non-native speaker in L1,the non-native accent's degree may vary within the same native language[4,5].

Accent variability is crucial for the systems of Automatic Speech Recognition (ASR), particularly with non-native speakers whose pronunciation may differ from that in standard training data. Traditional acoustic language models developed for standard language datasets do not meet the requirements for the recognition of accented speech. Addressing the challenge of recognizing accented speech by incorporating additional pronunciations and integrating samples into the training dataset is undesirable. This approach increases processing times and introduces additional noise, thus decreasing performance [6].

Despite the widespread deployment of ASR in many applications such as customer service, voice assistants, and accessibility tools, accents increase word error rates and diminish transcription accuracy [7], [8]. So, ASR systems must be improved to more effectively manage varied accents, making accent classification a crucial research priority for enhancing speech recognition performance.

Earlier methodologies for accent classification primarily depended on traditional machine learning methods before the emergence of deep learning, including Support Vector Machines (SVM) [9]. Hidden Markov Models (HMMs) [10-12], were employed. Many studies, including those by [10], have shown successful accent classification employing methods such as SVM, paired SVMs, and k-Nearest Neighbors (KNN), respectively. These approaches were constrained in their capacity to process data in its raw form. These methods are employed to extract features from audio recordings, relying heavily on handcrafted features (e.g., MFCCs and prosodic features such as energy, spectral tilt, and pitch), and subsequently utilizing those features to classify the audio into various categories [13].

In recent years, deep learning (DL) technologies have greatly improved speech-based applications. DL is a robust subset of machine learning that eliminates the need for manual feature extraction, as models acquire hierarchical representations directly from raw input data. Deep learning algorithms can extract features that enhance accent classification capabilities automatically [14], [15]. A substantial amount of processing power and data are essential for the DL's application to attain results with high accuracy[16]. Deep architectures generally surpass the ability of traditional techniques to capture complex, hierarchically structured statistical patterns in input[17].

Regarding native and non-native languages, English ranks as one of the most commonly spoken languages, and diverse English accents are naturally articulated across various societies. The principal varieties of English have interacted with different languages at several historical periods, experienced phonological alterations, and diverged from one another [18].

This research suggested developing a deep learning system for classifying regional UK English accents, which poses higher challenges than non-native accent classification due to the nuanced phonetic and prosodic variances among local accents. This work supports the development of more inclusive ASR systems and more personalized computer-assisted pronunciation training (CAPT) solutions.

The rest of the paper is organized as follows: Section 2 describes the related works. The methods and materials are described in Section 3. The analysis and the experimental results are presented in Section 4. And finally, the conclusions and future work are discussed in Section 5.

2. Related Works

This section examines the latest research on English accent classification, with a focus on feature representation, architectural advancements, accents, and datasets employed, which are outlined in **Table 1**.

The author in [19] presents a combination of an attention mechanism with a 1D CNN and a Bidirectional Gated Recurrent Unit (CNN-BiGRU) architecture is presented to classify six English speech accents. Choose the samples from the VoxForge dataset. MFCC and filter bank (FBank), two important acoustic features, are employed. The proposed model, 1D CNN-BiGRU-Attention, produced an average score of 85.52% for F1.

The authors in [18] utilized a transfer learning approach with a pre-trained AlexNet model to improve the UK-Ireland English Dialect Speech Dataset (UIED) performance. This study transforms audio signals into spectrogram images, enabling CNNs to handle two-dimensional data efficiently. The system achieved 93.38% accuracy in gender-dependent evaluations and 92.92% in gender-independent evaluations.

In[20], the study aspects at the effects of different spectral and time-frequency features, such as spectral roll-off, spectrograms, spectral centroid, chromatograms, and MFCC. The authors

showed that using linear-scale amplitude Mel-spectrograms significantly improves classification. The model's accuracy was 96.4% to 98.7% in classifying nine European accents. This was the best result from the dataset archive of accent speech, which kept pauses and used these features. However, the dataset had class imbalance because some accents had significantly fewer recordings than others. To correct this, the study used data augmentation.

The authors in [7] improved accent classification models by replacing old features like MFCCs and Mel-Spectrograms with the Hilbert Mel-Spectrogram, which is better at changing qualities and capturing speech's complex. This work used a 4-layer CNN model and chose audio samples from the Speech Accent Archive dataset. It can be seen from the results that the Hilbert Mel-spectrogram outperforms the traditional Fourier-based features when used to classify two accents, achieving an accuracy of up to 88%. Further comprehensive research is required to validate its efficacy. This feature has the potential to outperform the Fourier-based features.

The authors in[21] suggested a new Foreign Accent Identification (FAID) method called Multi-Kernel Extreme Learning Machine (MKELM) with a pairwise weighted strategy for classification of multi-class, utilizing MFCCs and prosodic features (such as pitch and energy) as input. The audio files are chosen from the speech accent archive dataset. The suggested approach outperforms other methods like ANN, SVM, LSTM, and ELM, yielding an accuracy of 84.72%.

The authors in[22] examine the effectiveness of using a method of filter-based feature selection to enhance accent classification. The methodology employs Mel-spectrograms as input features, derived from two pre-trained transfer learning models: InceptionV3 and MobileNetV2. The experiments demonstrate that the performance improved by using two filter-based feature selection methods, including mutual information (MI) and correlation analysis, to detect the top m% and n% most informative features, which are then combined. The finalized optimized feature set is provided to the SVM classifier for accent prediction. The suggested pipeline is evaluated using three public datasets—Speech Accent Archive, AccentDB, and the UK & Ireland English Dialect Speech Dataset—and achieves superior performance compared to other existing accent classification approaches reported in the literature.

The authors in[23] proposed three new deep learning models—Multi-task Pyramid Split Attention-DenseNet (MPSA-DenseNet), Pyramid Split Attention-DenseNet (PSA-DenseNet), and Multi-task DenseNet (Multi-DenseNet) for English accent classification. These models integrate multi-task learning and attention mechanisms with DenseNet to enhance model performance. Multi-task learning allows the models to handle multiple related tasks simultaneously, reducing overfitting and improving generalization. This study uses a common voice dataset and selects English accents from native speakers (USA, England) and non-native speakers (India, Germany, Hong Kong). MPSA-DenseNet surpasses other models, achieving state-of-the-art accuracy by combining DenseNet's feature reuse capabilities with attention mechanisms and multi-task learning, leading to improved accent classification across different regions.

Recent work shows significant advancements in accent classification using traditional and deep learning methods. Traditional methods like k-NN still perform well on small, balanced datasets. However, deep learning—especially CNN outperforms traditional approaches when sufficient data is available. Regional English accents—especially within the UK—pose additional challenges due to subtle phonetic variations and have received limited attention in the literature.

This study presents a lightweight 1D-CNN-GRU architecture designed to address this gap, which is intended for sequential acoustic features (MFCCs) to distinguish multiple English accents, including distinct UK and Irish, enabling high accuracy with low computational cost, in contrast to prior research that mostly relied on 2D spectrograms or transfer learning from image-based models.

Table 1. Overview of the accents' number, classification model, features, dataset, and performance for various prior research and the proposed study.

Ref	Year	Accents	Model	Features	Dataset	Results
[7]	2023	2	CNN	Hilbert Mel-Spectrogram	SAA	88%
[18]	2022	6	CNN	Spectrogram	UIED	92.92%
[20]	2022	9	CNN	Amplitude Mel-Spectrogram on a linear scale	SAA	96.4% to 98.7%
[21]	2024	6	MKELM	MFCCs+ Prosodic Features	SAA	84.72%
[22]	2024	6	MobileNetV2+ SVM	Mel Spectrogram	UIED	86.11%
[23]	2025	5	MPSA-denseNet models	MFCCs	Common Voice datasets	92.5%

3. Materials and Methods

This section presents the data collection, preprocessing, and feature extraction methods and then describes the proposed deep learning models. Additionally, the evaluation metrics employed to assess model performance.

3.1. Description of Dataset

This paper employs the UK and Ireland English Dialect Speech Dataset (UIED). This dataset comprises high-quality audio recordings of English utterances captured by individuals articulating various language dialects and is in the public domain. All audio in this dataset was recorded by volunteers who identified themselves as native speakers from the respective regions. This collection includes recordings of English from different dialects, with both male and female speakers from the United Kingdom and Ireland. The dataset was created for both linguistic analysis and speech technologies. There are 17,877 high-quality audio recordings (48 kHz, 16-bit, mono, Wave audio) recorded in a quiet environment. The recordings are usually 6.3 seconds long, but the longest is 20.1 seconds, and the shortest is 1.62 seconds. The average duration of each audio segment is 6 to 7 seconds. This dataset has six British Isles accents: Midlands, Southern England, Welsh English, Northern England, Irish English, and Scottish English. **Table 2** shows the dataset's class-wise distribution [24]**.**

Table 2. Distribution of the dataset considered for evaluating the proposed model's performance.

Accent label	Accent	Number of samples
1	Irish	450
2	Northern	2847
3	Southern	8492
4	Scottish	2543
5	Midlands	696
6	Welsh	2849

3.2 Standardizing Voice Files

To ensure the speech data were suitable for consistent feature extraction, an audio pre-processing step was designed and applied. All recordings were normalized to a fixed duration of six seconds. Longer recordings were truncated, whereas shorter ones were padded with zeros to achieve uniform length.

3.3 Feature Extraction

This study utilizes Mel-Frequency Cepstral Coefficients (MFCCs), which are recognized as highly effective features in speech-related [25]**.** The MFCCs were computed using Python-based tools for speech analysis [26]**.**

The speech signal is an important means of communication and carries a wide range of information. It can reflect vocabulary, emotion, intention, accent, dialect, age, speaker identity, gender, and other characteristics. However, creating systems that can accurately extract such information from speech still depends on using strong and reliable feature-extraction methods[27]. Speech signals are often analyzed over short intervals under the assumption that their

characteristics remain stable within those time spans. This method, known as short-term analysis, enables meaningful feature extraction[14]. Accordingly, each audio signal was divided into overlapping frames to capture temporal variations. The frame shift—the overlap between successive frames—was set to either one-half or one-third of the frame length to maintain overlap. To smooth the signal at the edges of each frame, a Hanning window was applied [19]. In this study, each frame was set to 30 milliseconds in length, with a 15 millisecond shift between frames. The extraction process began with the application of the Discrete Fourier Transform (DFT) to obtain the signal's frequency components. The Fast Fourier Transform (FFT) is an efficient approach for calculating the Discrete Fourier Transform (DFT). It decreases the complexity of calculations. This spectrum was then passed through a set of triangular filters distributed along the mel scale, designed to mirror the human ear's frequency sensitivity. The resulting mel spectrogram, usually plotted on either a linear or log scale, was then transformed using the Discrete Cosine Transform (DCT). This final step reduces feature redundancy and yields a compact representation of the original spectrum. The outcome is a set of MFCCs that effectively encode short-term spectral features on a perceptually relevant frequency scale [27], [28].

3.4 Feature Scaling

Audio signals collected from different environments and devices often have substantial variability, mainly in recording volume and background noise. Such inconsistencies, especially common in crowd-sourced datasets, can significantly negatively affect model performance [29]. To address this, feature scaling is essential for decreasing distortions and normalizing the audio signal into a specific range by utilizing z-normalization (z-score) and standardizing the data by removing the mean and by dividing it by the standard deviation of each recording. It is calculated from the following equation [20]:

$$x` = \frac{x - \mu}{\sigma} \quad (1)$$

Where x` is the normalized (standardized) value after z-normalization, x is the original feature value, μ denotes the mean value, and σ denotes the standard deviation.

3.5 Classification Model

3.5.1 1-D-CNN-GRU Model

Convolutional neural networks (CNNs) are deep feedforward neural networks that employ convolutional operations and a layered structure [30]. CNNs are among the most common methods in deep learning. They are highly efficient in feature extraction and excel with image and audio signal inputs [31-33]. CNN has excellent capability in extracting local features from speech signals; it comprises multiple layers, such as: pooling, convolutional and fully connected [26]. CNN models are functional in two stages first, feature extraction is performed, followed by classification of the extracted features using a fully connected layer [34]. The convolutional layer is the most essential processing unit of CNNs, responsible for most computations [35]. This layer employs convolution kernels to extract diverse feature maps, including lines, edges, and corners. Each filter utilizes the identical kernel to extract local properties from the input. Various filters with distinct weight vectors traverse the height and width of the input. Extracting several features is considered at each location [36]. Taking the kernel as k and the layer as l, we get the feature value $z^{l}_{i,i,k}$ at location (i, j). The representation of the function (i, j) can be mathematically expressed as demonstrated below [35]:

$$z^{l}_{i,i,k} = w^{l^T}_{k} \; x^{l}_{i,j} + b^{l}_{k} \quad (2)$$

Where $x^{l}_{i,j}$ is the input centered at the position (i, j) of the *lth* layer, $w^{l}_{i,j}$ and b^{l}_{k} represent the bias term of the *kth* kernel, *lth* layer, and the weight vector, respectively.

In CNNs, nonlinearity is commonly introduced through the non-saturating activation function known as ReLU. The ReLU function can be represented as shown by the provided equation below:

$$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & \text{Otherwise} \end{cases} \quad (3)$$

In this work, a MaxPooling layer was applied after each convolution to reduce the temporal dimension and retain dominant features[26]. Batch Normalization was constructed [20]. It standardizes each layer's inputs through training, improving convergence and stabilizing learning dynamics. This approach is generally employed in hidden layers and functions on mini-batches of data rather than individual samples, facilitating more stable gradient updates. The error gradient for a mini-batch is computed as:

$$\frac{1}{m}=\sum_{i=1}^{m}\frac{\partial\, l\,(x_i,\Theta)}{\partial\,\Theta} \quad \textbf{(4)}$$

Where m is the mini-batch size, Θ is the function of error minimization, and it indicates the input values of the dataset and the estimated error gradient for the complete dataset.

With sequential data, CNN performs poorly. To enhance the efficiency of the classification of English language accents, CNN is integrated with the gated recurrent unit (GRU), which is an upgraded algorithm for extracting global features. It efficiently handles the gradient vanishing issue of RNNs while highlighting benefits such as simple construction and excellent accuracy [37]. The GRU has two units: the first is the reset gate r_t and the second is the update gate z_t. It is assumed that the current input is r_t. GRU reregulates the rate at which information from the hidden layer, can be forgotten at a later time; it can be expressed as follows:

$$r_t=\sigma(W_r\,.[\,h_{t-1},\,x_t]) \quad \textbf{(5)}$$

Subsequently, z_t is utilized to determine how much of which preparatory hidden layer information $\tilde{h}_t$ can be preserved; it is expressed as follows:

$$z_t=\sigma(W_z\,.[\,h_{t-1},\,x_t]) \quad \textbf{(6)}$$

$$\tilde{h}_t=tanh(W_h.[\,h_{t-1}\;r_t,\,x_t]) \quad \textbf{(7)}$$

Where σ an activation function and W is a weight matrix.

At last, the output of the GRU consists of two components, which can be represented as follows:

$$h_t=(1-z_t)\,\tilde{h}_t+z_t h_{t-1} \quad \textbf{(8)}$$

To extract local temporal features from the input, the 1-D-CNN-GRU model begins with two 1D convolutional layers (32 and 64 filters) with max pooling, batch normalization, ReLU activation, and dropout (0.3 and 0.5). The output is then passed through two stacked GRU layers—128 units (returning sequences) and 64 units (returning the final state) to capture dependencies for the long-term dependencies. A dense layer with 128 units and ReLU activation, followed by dropout (0.6), refines the learned representation, and a final softmax layer outputs class probabilities across the target labels. **Figure 1** shows the entire pipeline of the proposed system. **Figure 2** illustrates the structure of the 1-D-CNN-GRU model.

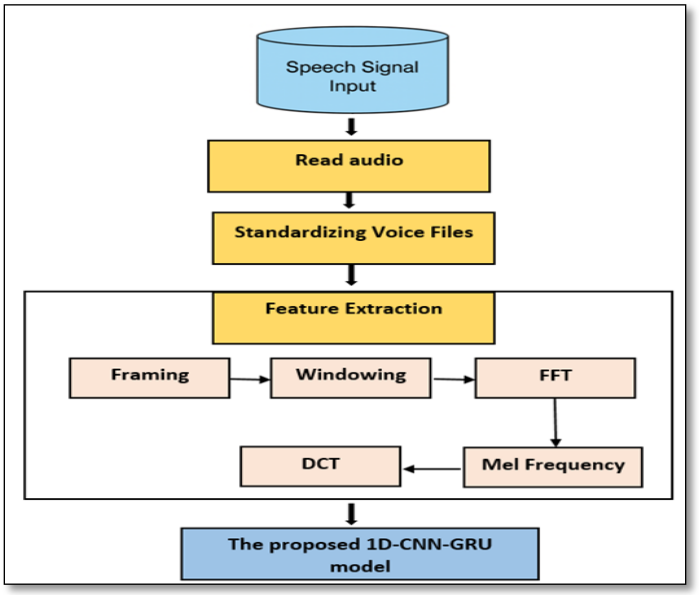

Figure 1. The overall pipeline of the proposed system

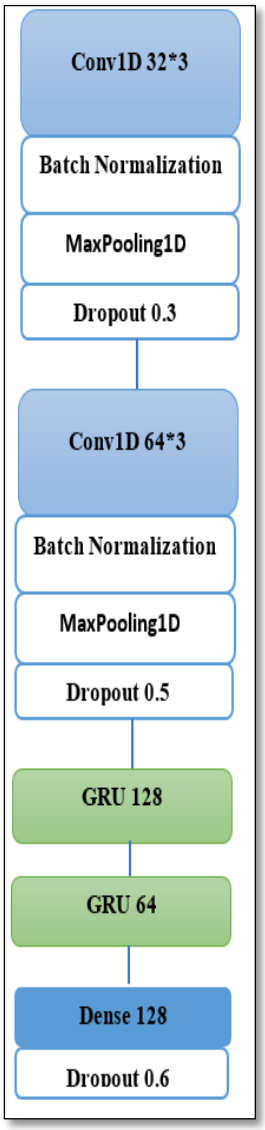

Figure 2. The structure of the proposed 1-D-CNN-GRU model.

3.6 Implementation Details

The models were initially trained using 60% of the training set's data, while 20% was used as validation. The experiments were executed and implemented in Python 3.11.6 [38] using Tensor Flow 2.18.0 [39] on a Windows system. The Adam optimizer was used to train the model with early stopping (patience = 5) based on validation loss and a learning rate of 0.001. Training was performed with a batch size of 32 and for up to 60 epochs. After the best model was selected on the validation set, the original training and validation splits were combined (80% of the data), from which a small stratified hold-out (~6.25% of that 80%, i.e., ~5% of the full dataset) was reserved for monitoring and fine-tuning that best model on the combined 80% split with a small (~5% of full) stratified holdout using a lower learning rate (1e-4). Final results were computed on the untouched 20% test set.

3.7 Evaluation Metrics

Several metrics were utilized to evaluate the classifier's performance. Accuracy, recall, precision, and F-measure were the primary measures utilized for accent classification. The performance metrics are explained below.

- **Accuracy** is computed by the division of the number of accurately predicted cases by the total predictions generated by the model, as shown by the equation below: [40]

- Accuracy $= \frac{TP + TN}{TP + TN + FP + FN}$ **(9)**

  True Negative (TN) indicates cases where both values are negative. True Positive (TP) represents the cases where the predicted values and the actual values are both positive. False Positive (FP) involves negative actual values incorrectly predicted as positive, while False Negative (FN) refers to positive actual values predicted as negative [41].
- **Recall:** Recall is the proportion of positive instances when both the expected and actual labels are correct, specifically the ratio of true positives to the total number of positive cases. It is sometimes referred to as the true positive rate. Recall can be computed as follows [41]:

  $\text{Recall} = \frac{TP}{TP+FN}$ **(10)**
- **Precision:** Precision is the measure that calculates the ratio of accurately positive and classified instances to the total number of positive instances identified by the system. It can be determined as below [40]:

  $\text{Precision} = \frac{TP}{TP+FP}$ **(11)**
- **Measure:** The F-measure is the weighted harmonic mean of recall and precision when false positives and false negatives are perfectly balanced. The F-measure formula is illustrated below [40]:
  - $\text{F-measure} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ **(12)**

4. Results and Discussion

This section examines and compares the proposed model with some existing accent classification methods to justify the effectiveness of this proposed method.

As presented in **Table 3.** the proposed 1D-CNN-GRU model outperforms other proposed models in this study, showing higher results when compared with other methods in the literature, achieving 98.71% accuracy and an F-measure 98.73% versus 92.92% accuracy from a CNN with spectrogram features [18], and 86.11% from MobileNetV2 with SVM on Mel spectrograms [22]. Highlighting its superior effectiveness on the UIED dataset**. Table 4.** provides a comparative overview of the proposed 1D-CNN-GRU model against recent studies, emphasizing its advantages and current limitations.

Table 3. Classification accuracy obtained by the proposed methods on the UIED dataset.

Ref	Accents	Model	Features	Dataset	Results
[18]	6	CNN	Spectrogram	UIED	92.92%
[22]	6	MobileNetV2+ SVM	Mel Spectrogram	UIED	86.11%
proposed model	6	1D-CNN–GRU	MFCC	UIED	98.71 %

Table 4. Comparison of advantages and drawbacks between the proposed model and previous studies.

Study / Model	Key Advantages	Drawbacks / Limitations
[18]	• Utilizes image-based CNN feature extraction • Captures rich spectral information.	• Computationally expensive • Moderate accuracy (92.92 %).
[22]	• Feature selection methods improved discriminative ability.	• Requires two-stage training (CNN + SVM) • Lower accuracy (86.11 %).
proposed model	• High accuracy (98.71 %) on UIED dataset • Lightweight and computationally efficient • Combines local (CNN) and temporal (GRU) feature learning • Robust representation of accent-specific patterns	• Limited to MFCC features • Requires further validation on spontaneous and noisy speech

5. Conclusion

In this research, a deep learning model based on a 1D-CNN-GRU architecture was designed and tested for the task of English accent classification. The experiments were carried out using the UIED dataset, and the proposed system showed noticeably stronger performance than the methods currently reported in the literature. On this dataset, the model achieved a peak accuracy of 98.71%, indicating that it can learn and distinguish the subtle acoustic cues that are tied to different English accents.

For future work, it would be beneficial to expand the dataset so that it includes a broader range of accents, more spontaneous speech, and a larger and more diverse pool of speakers, which would help improve the model's generalizability. It may also be useful to complement MFCCs with other acoustic representations, such as filter bank energies or prosodic features, in order to provide richer input information. Another direction worth exploring is the integration of transformer-based representations, such as Wav2Vec2, which may offer more discriminative, accent-specific embeddings and further improve the overall classification performance.

Acknowledgment

The authors express their gratitude to the Department of Computer Science staff, College of Science, at the University of Baghdad for their inspiration and support.

Conflict of Interest

The authors declare that they have no conflicts of interest.

Funding

No funding.

References

1. Habbash M, Mnasri S, Alghamdi M, Alrashidi MQ, Tarawneh AS, Gumair A, Hassanat ABA. Recognition of Arabic accents from English spoken speech using deep learning approach. IEEE Access. 2024;12:37219–37230. https://doi.org/10.1109/ACCESS.2024.3374768.
2. Grigaliūnaitė J, Melnik-Leroy GA. Automatic accent identification using less data: a shift from global to segmental accent. Arab J Sci Eng. 2025;50(10):7481–7494. https://doi.org/10.1007/s13369-024-09344-4.
3. O'Grady W, Archibald J, Aronoff M, Rees-Miller J. Contemporary Linguistics Analysis. 8th ed. 1992.
4. Flege JE, Schirru C, MacKay IRA. Interaction between the native and second language phonetic subsystems. Speech Commun. 2003;40(4):467–491. https://doi.org/10.1016/S0167-6393(02)00128-0.
5. Behravan H, Hautamäki V, Kinnunen T. Factors affecting i-vector based foreign accent recognition: a case study in spoken Finnish. Speech Commun. 2015;66:118–129. https://doi.org/10.1016/j.specom.2014.10.004.
6. Bogach N, et al. Speech processing for language learning: a practical approach to computer-assisted pronunciation teaching. Electronics. 2021;10(3):235.
7. Walsh D, Dev S, Nag A. Hilbert–Huang-transform based features for accent classification of non-native English speakers. In: 2023 34th Irish Signals and Systems Conf (ISSC); 2023. https://doi.org/10.1109/ISSC59246.2023.10162075.
8. Aju O. A review of accent-based automatic speech recognition models for e-learning environment. Covenant J ICT. 2022;10(2). Available from: https://journals.covenantuniversity.edu.ng/index.php/cjict/article/view/3146
9. Rizwan M, Anderson DV. A weighted accent classification using multiple words. Neurocomputing. 2018;277:120–128. https://doi.org/10.1016/j.neucom.2017.01.116.
10. Tang H, Ghorbani AA. Accent classification using support vector machine and hidden Markov model. In: Lect Notes Comput Sci. Springer; 2003. p. 629–631. https://doi.org/10.1007/3-540-44886-1_65.
11. Hou J, Liu Y, Zheng TF, Olsen J, Tian J. Multi-layered features with SVM for Chinese accent identification. In: ICALIP 2010 Proc. IEEE; 2010. p. 25–30. https://doi.org/10.1109/ICALIP.2010.5685023.

12. Salifu A, Mensah HN, Tchao ET, Acheampong FA, Agbemenu AS, Kponyo JJ. Enhancing speech recognition through diverse shared features accent classification. Int J Speech Technol. 2025;28(2):461–481. https://doi.org/10.1007/s10772-025-10198-w.
13. Ibrahim NJ, Idris MYI, Yusoff MYZM, Rahman NNA, Dien MI. Robust feature extraction based on spectral and prosodic features for classical Arabic accents recognition. Malays J Comput Sci. 2019;(Spec Issue 3):46–72. https://doi.org/10.22452/mjcs.sp2019no3.4.
14. Jassim S, Abdulmohsin HA. Accent classification using machine learning techniques: a review. Int J Comput Inf Syst Ind Manag Appl. 2025;17:421–451.
15. Dar MA, Jagalingam P. Machine learning and deep learning approaches for accent recognition: a review. IEEE Access. 2025.
16. Ölmez E, Akdoğan V, Korkmaz M, Er O. Automatic segmentation of meniscus in multispectral MRI using regions with convolutional neural network (R-CNN). J Digit Imaging. 2020;33(4):916–929. https://doi.org/10.1007/s10278-020-00329-x.
17. Qiu J, Wu Q, Ding G, Xu Y, Feng S. A survey of machine learning for big data processing. EURASIP J Adv Signal Process. 2016;2016(1):1–16. https://doi.org/10.1186/s13634-016-0355-x.
18. Cetin O. Accent recognition using a spectrogram image feature-based convolutional neural network. Arab J Sci Eng. 2023;48(2):1973–1990. https://doi.org/10.1007/s13369-022-07086-9.
19. Ke W. Study on recognition and classification of English accents using deep learning algorithms. J Intell Syst. 2023;32(1). https://doi.org/10.1515/jisys-2023-0174.
20. Mikhailava V, Lesnichaia M, Bogach N, Lezhenin I, Blake J, Pyshkin E. Language accent detection with CNN using sparse data from a crowd-sourced speech archive. Mathematics. 2022;10(16). https://doi.org/10.3390/math10162913.
21. Kashif K, Alwan A, Wu Y, De Nardis L, Di Benedetto MG. MKELM based multi-classification model for foreign accent identification. Heliyon. 2024;10(16). https://doi.org/10.1016/j.heliyon.2024.e36460.
22. Bhadra R, Sahu M, Agrebi M, Singh PK, Badr Y. A hybrid deep feature selection framework for speaker accent recognition. In: Leveraging Computer Vision to Biometric Applications. Chapman and Hall/CRC; 2024. p. 154–177. https://doi.org/10.1201/9781032614663-8.
23. Song T, Nguyen LTH, Ta TV. MPSA-DenseNet: a novel deep learning model for English accent classification. Comput Speech Lang. 2025;89. https://doi.org/10.1016/j.csl.2024.101676.
24. Demirsahin I, Kjartansson O, Gutkin A, Rivera C. Opensource multispeaker corpora of the English accents in the British Isles. In: Proc 12th Int Conf Lang Resour Eval (LREC); 2020. p. 6532–6541. Available from: https://aclanthology.org/2020.lrec-1.804/
25. Ali AT, Abdullah H, Fadhil MN. Speaker recognition system based on Mel frequency cepstral coefficient and four features. Iraqi J Comput Commun Control Syst. 2021;1(4):8.
26. Hussien AAR, Abdullah NAZ. A review for Arabic sentiment analysis using deep learning. Iraqi J Sci. 2023;64(12):6572–6585. https://doi.org/10.24996/ijs.2023.64.12.37.
27. Mohammed SN, Hassan AK. Automatic voice activity detection using fuzzy-neuro classifier. J Eng Sci Technol. 2020;15(5):2854–2870.
28. Zheng F, Zhang G, Song Z. Comparison of different implementations of MFCC. J Comput Sci Technol. 2001;16(6):582–589. https://doi.org/10.1007/BF02943243.
29. Alashaikh AS, Alhazemi FM. Efficient mobile crowdsourcing for environmental noise monitoring. IEEE Access. 2022;10:77251–77262.
30. Al-Jumaili Z, Bassiouny T, Alanezi A, Khan W, Al-Jumeily D, Hussain AJ. Classification of spoken English accents using deep learning and speech analysis. In: Lect Notes Comput Sci. Springer; 2022. p. 277–287. https://doi.org/10.1007/978-3-031-13832-4_24.
31. Kumar R, Singh K, Mahato DP, Gupta U. Face-based age and gender classification using deep learning model. Procedia Comput Sci. 2024;235:2985–2995. https://doi.org/10.1016/j.procs.2024.04.282.
32. Abdulmohsin HA, Stephan JJ, Al-Khateeb B, Hasan SS. Speech age estimation using a ranking convolutional neural network. In: Lect Notes Netw Syst. Springer; 2022. p. 123–130. https://doi.org/10.1007/978-981-19-0604-6_11.
33. Ahmed HM, Mahmoud HH. Effect of successive convolution layers to detect gender. Iraqi J Sci. 2018;59(3):1717–1732. https://doi.org/10.24996/IJS.2018.59.3C.17.

34. Zare S, Ayati M. Simultaneous fault diagnosis of wind turbine using multichannel convolutional neural networks. ISA Trans. 2021;108:230–239. https://doi.org/10.1016/j.isatra.2020.08.021.
35. Ozer I. Pseudo-colored rate map representation for speech emotion recognition. Biomed Signal Process Control. 2021;66:102502. https://doi.org/10.1016/j.bspc.2021.102502.
36. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang G, Cai J, Chen T. Recent advance in convolutional neural networks. Pattern Recognit. 2018;77:354–377. https://doi.org/10.1016/j.patcog.2017.10.013.
37. Yevnin Y, Chorev S, Dukan I, Toledo Y. Short-term wave forecasts using gated recurrent unit model. Ocean Eng. 2023;268:113389. https://doi.org/10.1016/j.oceaneng.2022.113389.
38. Python Software Foundation. Python: version 3.11.6. 2023. Available from: https://www.python.org/downloads/release/python-3116/
39. TensorFlow Development Team. TensorFlow: version 2.18.0. 2024. Available from: https://pypi.org/project/tensorflow/2.18.0/
40. [40] Ismail M, Maarof MA, Hamzah FA, Jeffrey YM, Abidin AZ, Omar N, Awang S. Development of a regional voice dataset and speaker classification based on machine learning. J Big Data. 2021;8(1):1–18. https://doi.org/10.1186/s40537-021-00435-9.
41. Ozer I, Cetin O, Gorur K, Temurtas F. Improved machine learning performances with transfer learning to predicting need for hospitalization in arboviral infections against the small dataset. Neural Comput Appl. 2021;33(21):14975–14989. https://doi.org/10.1007/s00521-021-06133-0.