Research Article

Wenjuan Ke*

Study on recognition and classification of English accents using deep learning algorithms

https://doi.org/10.1515/jisys-2023-0174
received September 19, 2023; accepted December 05, 2023

Abstract: The recognition and classification of English accents have high practical value in areas such as security management and information retrieval. This study introduced two English accent features, filter bank (FBank) and Mel-frequency cepstral coefficient (MFCC), based on deep learning techniques. It then combined convolutional neural network (CNN), gated recurrent unit, and an attention mechanism to design a 1D CNN-BiGRU-Attention model for English accent recognition and classification. Experimental tests were conducted on the VoxForge dataset. The results showed that compared to MFCC, FBank performed better in English accent recognition and classification, and 70FBank achieved the highest *F*1 value. Among the recurrent neural network, long short-term memory, and other models, the BiGRU model had the best performance. The average *F*1 value of the 1D CNN-BiGRU-attention model was the highest, reaching 85.52%, and all the *F*1 values were above 80% for different accents, indicating that the addition of the attention mechanism effectively improved the model's recognition and classification effectiveness. The results prove the reliability of the method proposed in this article for English accent recognition and classification, making it suitable for practical application and promotion.

Keywords: deep learning, English accent, recognition and classification, FBank

AMS Mathematics Subject Classification number: 68T07

1 Introduction

Speech recognition technology can enhance human–computer interaction and has significant applications in various fields such as smart homes and medical rehabilitation [1]. In actual speech, the presence of accents can significantly impact recognition performance [2]. Accents refer to different pronunciations of words [3]. Within the same language, different accents can lead to variations in tone, duration, and other aspects [4]. Taking English accents as an example, American English tends to have more rolled sounds and a flat intonation, while British English avoids rolling the tongue and has great pitch variation. Indian English is spoken at a fast pace, and French English often substitutes “z” for “th.” Furthermore, even within the same country, different regions have their own distinct accents; however, all of these variations still fall under the umbrella of the English language. By recognizing and classifying different English accents, it is possible to provide corresponding services for customers in various self-service industries such as tourism and restaurants. This can also assist in determining the origin of individuals in areas of immigration control and crime investigation for security purposes. Deep learning algorithms are capable of automatically learning features from raw data, and they adopt an end-to-end learning approach, making model construction relatively simple. They possess

*** Corresponding author: Wenjuan Ke,** School of Foreign Languages, Hubei University of Science and Technology, Building 31, Tianjieshuhua City, Xian'an District, Xianning, Hubei 437100, China, e-mail: wj50890@163.com

excellent transfer learning capabilities and demonstrate outstanding performance in handling large-scale complex data, particularly in the field of computer vision, such as object detection and facial recognition [5]. These algorithms find wide applications in areas like autonomous driving, video surveillance, virtual reality, etc. [6]. In addition, deep learning has been successfully applied in the field of natural language processing, such as machine translation and sentiment analysis [7]. In the medical field, deep learning-based pathology image analysis and diagnosis contribute to improving diagnostic accuracy [8]. In speech recognition, deep learning is widely used in applications like voice assistants and speech-to-text conversion [9]. Various deep learning methods have been extensively researched. Zhang et al. [10] designed two models, multi-layer cellular neural network-connectionist temporal classification (MCNN-CTC) and SENet (SE)-MCNN-CTC, for Chinese speech recognition. Through experiments, they found that the relative error rate of SE-MCNN-CTC decreased by 13.51%, resulting in a final error rate decrease of 22.21%, indicating higher generalization performance. The study conducted by Manohar and Logashanmugam [11] focused on speech emotion recognition and proposed a hybrid deep learning approach, which was found to outperform other models in terms of performance. Kumar et al. [12] investigated the technology of converting speech into text for hearing-impaired students and utilized a deep learning-based model to extract features from audio and video, achieving a word error rate of 6.59%. The study conducted by Seki et al. [13] proposed a deep neural network that combines filter banks (fBanks). The experimental results showed a 5.8% reduction in word errors among ten utterances. This work designed a deep learning algorithm for English accent recognition and classification, and its reliability was demonstrated through experiments on different accent datasets. This work provides a new and reliable method for practical applications in areas such as public safety and language learning involving English accent recognition and classification. This study has taken into account the diversity of English accents, and the designed model demonstrates good performance in recognizing and classifying different accents. This study provides a new approach to address the issues of accent diversity and adaptability, which contributes to promoting wider and more effective applications of speech recognition classification technology.

2 Extraction of English accent features

As an acoustic signal, the English accent needs to undergo preprocessing and feature extraction before it can be recognized and classified using deep learning algorithms. First, the energy of the high-frequency part of the signal is enhanced through pre-emphasis, which is achieved by applying a first-order digital filter. The corresponding equation is:

$$H(z) = 1 - az^{-1}, \tag{1}$$

where a stands for the pre-emphasis coefficient.

Then, by utilizing the short-term stationary characteristics of the signal, it is divided into multiple speech frames, each typically having a length of 10–30 ms. The overlapping portion between adjacent frames, known as frame shift, is usually around half to one-third of the frame length. After framing, window functions are applied to maintain smoothness at both ends of each speech frame. Commonly used window functions are listed in Table 1.

Table 1: Commonly used window functions

Name	Equation	Feature
Rectangular window	$w(n) = \begin{cases} 1, 0 \le n \le L–1 \\ 0, \text{else} \end{cases}$	High sidelobes and great spectral leakage
Hanning window	$w(n) = \begin{cases} 0.5–0.5\cos[2\pi n/(L–1)],\ 0 \le n \le L–1 \\ 0, \text{else} \end{cases}$	Good suppression of spectral leakage, but low resolution
Hamming window	$w(n) = \begin{cases} 0.54–0.46\cos[2\pi n/(L–1)],\ 0 \le n \le L–1 \\ 0, \text{else} \end{cases}$	Slower sidelobe attenuation than Hanning window, better low-pass characteristics

According to Table 1, this study utilizes a Hamming window with a frame length of 25 ms and a frame shift of 10 ms in the processing of English accent signals. Finally, to preserve the effective segment of the signal, endpoint detection is performed using the commonly employed dual threshold method [14]. Let the signal obtained after framing and windowing be $x_i(m)$. The steps of endpoint detection are as follows.

(1) The energy per frame is calculated – $E_i = \sum_{m=1}^{n} x_i^2(m)$.

(2) A small value called σ is set for center clipping processing – $x_i(m) = \begin{cases} x_i(m), |x_i(m)| \geq \sigma \\ 0, |x_i(m)| < \sigma \end{cases}$

(3) The zero crossing rate per frame is calculated – $\text{ZCR} = \frac{1}{2}\sum_{m=1}^{N} |\text{sign}[x_i(m)] - \text{sign}[x_i(m-1)]|$, $\text{sign}[x_i(m)] = \begin{cases} 1, |x_i(m)| \geq 0 \\ -1, |x_i(m)| < 0 \end{cases}$.

(4) The starting and ending points of the dual threshold triggering are determined. The continuous signal between them are considered as valid signal.

After the preprocessing of English accent signals is completed, feature extraction can be performed. Commonly used features include linear predictive cepstral coefficients [15], Mel-frequency cepstral coefficients (MFCC) [16], and so on. Among them, MFCC and FBank are closer to how the human ear processes audio; therefore, this study mainly focuses on studying these two features. The extraction process is shown in Figure 1.

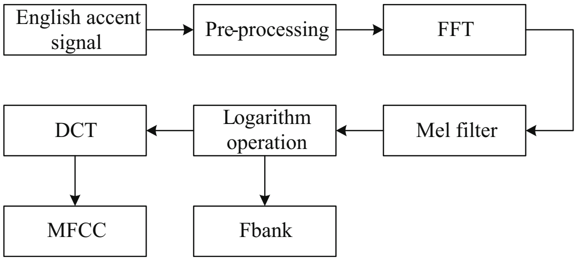

Figure 1: FBank and MFCC feature extraction.

According to Figure 1, after pre-emphasis and other preprocessing of the English oral signal, fast Fourier transform is performed on the obtained signal to complete time-frequency conversion.

$$X_t(k) = \sum_{n=0}^{K-1} x_t(n)\exp\left(-j\frac{2\pi nk}{K}\right), k = 0, 1, \cdots, K-1, \tag{2}$$

where $x_t(n)$ stands for the $x_t(n)$-th sampling point of the t-th frame of the signal.

Then, the power spectrum of the t-th frame signal is calculated as follows:

$$P_t(k) = |X_t(k)|^2/K, k = 1, 2, \cdots, K/2 + 1. \tag{3}$$

Actual frequency f is mapped to Mel frequency f_{mel} as follows:

$$f_{\text{mel}} = 2{,}595 \times \lg(1 + f/700). \tag{4}$$

Mel filters are constructed. The frequency response obtained by the m-th filter is as follows:

$$H_m(\delta) = \begin{cases} 0, \delta < f_L(m), \\ \dfrac{\delta - f_C(m)}{f_C(m) - f_L(m)}, f_L(m) < \delta \leq f_C(m), \\ \dfrac{f_C(m) - \delta}{f_H(m) - f_C(m)}, f_C(m) < \delta \leq f_H(m), \\ 0, \delta > f_H(m), \end{cases} \tag{5}$$

where $f_C(m)$ refers to the central frequency of the filter, and $f_L(m)$ and $f_H(m)$ are upper and lower cutoff frequencies, respectively.

The signal is passed through a group of Mel filters, and the logarithmic energy output of each filter can be written as follows:

$$S_t(m) = \ln\left(\sum_{K=0}^{N-1} |X_t(k)|^2 H_m(\delta)\right), \tag{6}$$

where $S_t(m)$ is the m-th dimension feature of FBank extracted from the t-th frame signal.

Discrete cosine transform (DCT) is performed on $S_t(m)$ to obtain

$$C_t(n) = \sum_{n=1}^{N} S_t(m)\cos\left[\frac{\pi n}{M}(m - 0.5)\right], n = 1, 2, \cdots, L, \tag{7}$$

where $C_t(n)$ is the MFCC feature, L is the number of orders, and M is the number of filters.

3 Recognition and classification using deep learning algorithms

Convolutional neural networks (CNNs) are commonly used algorithms in deep learning [17], with wide applications in various areas such as image recognition [18]. They perform excellently at extracting local features from data, which is why this study chose CNN to extract local features from English accent signals. In actual operations, a window with a size of k slides over an input feature map to complete the convolution. It is assumed that the input feature map of the l − 1-th layer is x^{l-1}. The output is obtained by multiplying x^{l-1} with the convolutional kernel k^l of the l-th layer, summing them up, adding the bias term b^l of the current layer, and applying an activation function. For the j-th convolutional kernel, the convolution operation can be written as follows:

$$x_j^l = f\left(\sum_{i\in M_j} x_i^{l-1} \otimes k_{ij}^l + b_j^l\right), \tag{8}$$

where M_j refers to the height of the j-th convolutional kernel.

Pooling operations aim to reduce the complexity of calculation to accelerate network convergence. First, the j-th feature map in the l − 1-th layer, x_j^{l-1}, is downsampled. Then, it is multiplied with β_j^l, the j-th parameter of the l-th layer. Then, the output and the bias b_j^l of the current layer are added together. After the operation using the activation function, x_j^l is output. The pooling operation is written as follows:

$$x_j^l = f[\beta_j^l \text{downsample}(x_j^{l-1}) + b_j^l], \tag{9}$$

where downsample refers to to downsampling operation.

However, CNN has poor performance in handling sequential data. To further improve the effectiveness of English accent recognition classification, CNN is combined with a recurrent neural network (RNN). Within the RNN, the gated recurrent unit (GRU) is an improved algorithm [19] that effectively alleviates the gradient vanishing problem of RNN while also possessing advantages such as a simple structure and high accuracy. GRU has two units: update gate z_t and reset gate r_t. It is assumed that the current input is r_t. GRU determines how much hidden layer information h_{t-1} can be forgotten at the late moment through r_t, which can be written as follows:

$$r_t = \sigma(W_r\cdot[h_{t-1}, x_t]). \tag{10}$$

Then, z_t is used to determine how much h_{t-1} and preparatory hidden layer information $\widetilde{h}_t$can be reserved, which can be written as follows:

$$z_t = \sigma(W_z\cdot[h_{t-1}, x_t]), \tag{11}$$

$$\widetilde{h}_t = \tanh(W_h\cdot[h_{t-1}r_t, x_t]). \tag{12}$$

Finally, the output h_t of GRU is composed of two parts, which can be written as follows:

$$h_t = (1 - z_t)\tilde{h}_t + z_t h_{t-1}, \tag{13}$$

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

In order to capture information from both past and future moments, this article adopts BiGRU, and the calculation formula can be written as follows:

$$\overrightarrow{h_t} = \mathrm{GRU}(\overrightarrow{h_{t-1}}, x_t), \tag{14}$$

$$\overleftarrow{h_t} = \mathrm{GRU}(\overleftarrow{h_{t-1}}, x_t), \tag{15}$$

$$h_t = W_t\overrightarrow{h_t} + V_t\overleftarrow{h_t}, \tag{16}$$

where $\overrightarrow{h_t}$ represents the output of forward GRU, $\overleftarrow{h_t}$ represents the output of reverse GRU, and W_t and V_t are weight matrices.

In the recognition and classification of English accent, special attention needs to be paid to the unique pronunciation styles of different accents. The attention mechanism [20] can adjust weight coefficients to modify the importance of features. Therefore, this study combines the attention mechanism with CNN-BiGRU. Local features are extracted from the extracted English speech signal features mentioned earlier using a two-layer 1D CNN. By adopting a two-layer structure, the features extracted by the first convolutional layer can be passed on to the second layer, enabling further learning of more complex and abstract features, thus achieving deep-level feature learning. BiGRU is used for extracting global features. An attention module is then added to adjust feature weights. In order to convert the model's output into probabilities for each category, a softmax function is used in the last layer for classification and recognition of different English accents. The designed model structure is shown in Figure 2.

As shown in Figure 2, the output sequence of the BiGRU is denoted as s_t. The target attention weight can be written as follows:

$$a_t = \tanh(s_t). \tag{17}$$

The class probability vector generated by softmax is:

$$p_t = [\exp(a_t)]/\left[\sum_{t=1}^{m}\exp(a_t)\right]. \tag{18}$$

Finally, the calculation formula of the probability distribution of different classes is as follows:

$$P = \mathrm{softmax}(w_a v + b_a), \tag{19}$$

where w_a and b_a are the weight and bias, and v is the weighting vector of a_t, $v = \sum_{t=1}^{m} a_t p_t$.

4 Results and analysis

4.1 Experimental setting

The experimental dataset was VoxForge [21], which includes processed data of various English accents. From this dataset, six accents with a relatively large amount of data were selected, as listed in Table 2.

80% of the data was used for training, while the remaining 20% was used for testing. The model was built using the Keras framework and programmed in Python 3.6 on an Ubuntu operating system. The learning rate for the 1D CNN-BiGRU-Attention model was set to 0.001, the batch size was set to 64, and the number of training epochs was set to 120. The Adam optimizer was used. Evaluation of the model's recognition and classification performance was based on a confusion matrix (Table 3).

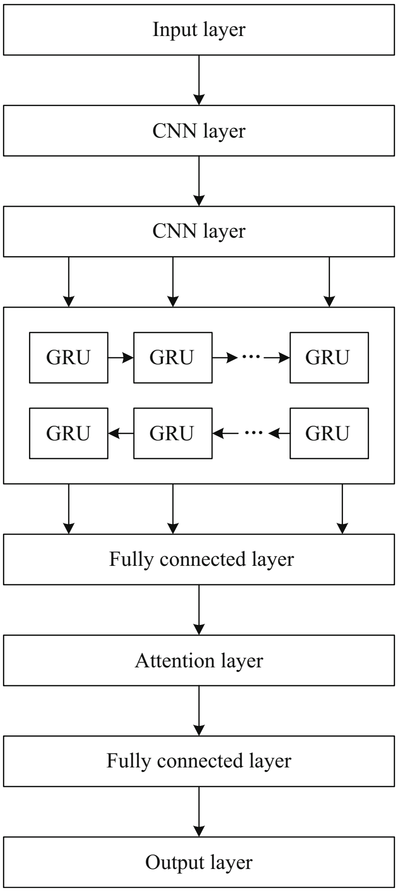

Figure 2: 1D CNN-BiGRU-attention model.

Table 2: Experimental dataset

Accent	Number
American English (AM)	8,306
British English (BR)	8,134
European English (EU)	7,842
Canadian English (CA)	3,405
Indian English (IN)	2,407
Australian English (AU)	2,323

Table 3: Confusion matrix

	Recognition and classification result	
Real class	Positive case	Negative case
Positive case	TP	FN
Negative case	FP	TN

(1) Precision: P = TP/(TP + FP)
(2) Recall rate: R = TP/(TP + FN)
(3) $F1$ value: $F1 = 2PR/(P + R)$

4.2 Result analysis

The commonly used MFCC feature in the extraction of English accent features was 39-dimensional, consisting of 12 dimensions for output, 1 dimension for logarithmic energy, and first-order and second-order differences. The default FBank group for FBank was set to 23. First, the differences between using FBank features and MFCC features in a 1D CNN-BiGRU-Attention model were compared. The FBank groups for FBank were adjusted to 23, 45, 70, 95, and 115. The recognition classification results are presented in Figure 3.

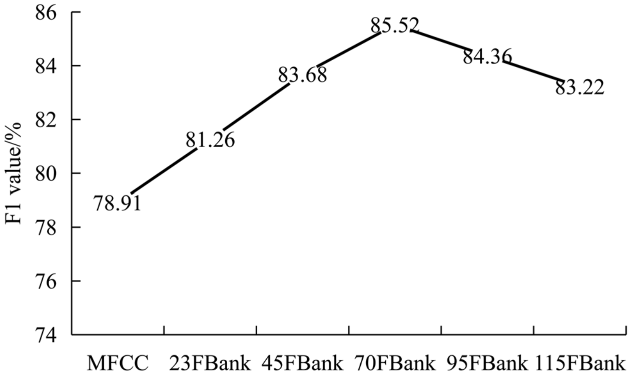

Figure 3: Influence of English accent features on model identification and classification results.

From Figure 3, it can be observed that when using the MFCC as the feature, the $F1$ value of the model in English accent recognition and classification was 78.91%. However, there was a significant improvement in $F1$ value when using the FBank as the feature. When the number of FBanks was set to 23, the $F1$ value of the FBank reached 81.26%, representing an increase of 2.35% compared to MFCC. Compared to the FBank, the MFCC lost some correlation details after DCT processing, resulting in inferior performance in recognition classification. As the number of FBanks increased on the FBank features, the model's $F1$ value initially increased and then decreased. Comparatively speaking, the highest $F1$ value was achieved with 70FBank, which was 6.61% higher than the MFCC. This may be because the continuous increase in FBanks can lead to an increase in signal noise, making the training process more complex and resulting in a poorer recognition and classification effect. Based on this, 70FBank was used as the feature in subsequent experiments.

In the case of fixing the CNN layer and attention layer, the influence of different RNNs on the model performance was compared, as shown in Table 4.

Table 4: Influence of different RNNs on model performance

	Precision (%)	Recall rate (%)	*F*1 value (%)
1D CNN-RNN-Attention	71.26	68.77	69.99
1D CNN-LSTM-Attention	75.44	73.11	74.26
1D CNN-BiLSTM-Attention	78.21	74.25	76.18
1D CNN-GRU-Attention	84.33	81.26	82.77
1D CNN-BiGRU-Attention	86.72	84.36	85.52

The different choices of RNN had a significant impact on the recognition and classification results of the model, as observed from Table 4. When using the RNN, the *F*1 score for English accent recognition was only 69.99%, which was the lowest. After replacing RNN with the long short-term memory (LSTM), the *F*1 value increased to 74.26%, showing an improvement of 4.27% compared to the RNN. Further replacing the RNN with the GRU led to an even higher *F*1 score of 82.77%, indicating an improvement of 12.78% compared to the RNN and an additional improvement of 8.51% compared to the LSTM. These results indicated that the GRU performed the best. Comparing unidirectional and bidirectional RNNs, the BiLSTM had an *F*1 value of 76.18%, which was a 1.92% improvement over the LSTM, while the BiGRU had an *F*1 value of 85.52%, which was a 2.75% improvement over GRU. This suggested that models using bidirectional networks could learn more features and improve its performance in English accent identification and classification.

Table 5: Influence of the attention mechanism on the model performance

	Precision (%)	Recall rate (%)	*F*1 value (%)
1D CNN-BiLSTM	76.01	74.31	75.15
1D CNN-BiLSTM-Attention	78.21	74.25	76.18
1D CNN-BiGRU	85.22	82.41	83.79
1D CNN-BiGRU-Attention	86.72	84.36	85.52

The influence of the attention mechanism on the model performance was compared, and the results are shown in Table 5.

The presence or absence of the attention mechanism also had a certain impact on the identification and classification results of the model, as observed from Table 5. Without adding the attention mechanism, the *F*1 value of the 1D CNN-BiLSTM approach was 75.15%, which decreased by 1.03% compared to the 1D CNN-BiLSTM-Attention approach. Similarly, the *F*1 value of the 1D CNN-BiGRU approach was 83.79%, which decreased by 1.73% compared to the 1D CNN-BiGRU-Attention approach. These results indicated that the ability to classify weights using the attention mechanism could further enhance the performance of the model in English accent recognition, thereby demonstrating the effectiveness of the method proposed in this work.

Table 6: *F*1 value of different approaches for the identification and classification of different English accents (unit: %)

	AM	BR	EU	CA	IN	AU
1D CNN-LSTM-Attention	77.33	78.46	76.59	60.12	80.76	72.30
1D CNN-BiLSTM-Attention	81.26	83.54	70.12	65.42	81.26	75.48
1D CNN-GRU-Attention	75.32	91.27	87.26	75.61	81.22	85.94
1D CNN-BiGRU-Attention	92.36	92.16	80.33	80.12	87.33	80.82

Finally, the *F*1 value of different approaches for the identification and classification of different English accents was compared, and the results are presented in Table 6.

From Table 6, it can be seen that first, different methods achieved good recognition and classification results on Indian accent (IN), with *F*1 values above 80%, which may be because IN was significantly different from the other accents and had obvious features. Comparing the different approaches, the 1D CNN-LSTM-Attention approach achieved the highest *F*1 value of 80.76% on IN but performed the worst in recognizing Canadian English (CA) with a lowest *F*1 value of only 60.12%. On the other hand, the 1D CNN-BiLSTM-Attention approach showed the best recognition and classification performance on BR but performed poorly in recognizing Australian English (AU), which may be because the small sample size of AU. As for British English (BR), the 1D CNN-GRU-Attention approach reached an *F*1 value of 91.27%; for American English (AM) and BR, their *F*1 values were above 90%; however, it had the poorest performance in recognizing CA with an *F*1 value of only 80.12%. Different methods showed different results in accent recognition and classification tasks, indicating that these methods had varying effects on feature extraction and classification. However, overall, the 1D CNN-BiGRU-Attention approach had an *F*1 value of over 80% for each type of accent, with the highest average *F*1 value reaching 85.52%, which demonstrated the reliability of this method.

5 Conclusion

This article proposed a 1D CNN-BiGRU-Attention model based on deep learning for recognizing and classifying different English accents. The experiments showed that FBank, as a feature, exhibited better recognition and classification performance compared to MFCC. In comparison with other methods, the designed 1D CNN-BiGRU-Attention model demonstrated excellent performance in recognizing and classifying different accents, achieving a superior average *F*1 score of 85.52%. Therefore, the proposed method is feasible and can be applied in practice. However, this study also has some limitations. For instance, in terms of features, only the performance of MFCC and FBank were compared. Additionally, the selection of models did not consider deeper and more complex CNN structures. Therefore, future work needs to delve deeper into feature research to explore the performance of different features in English accent recognition and classification. Furthermore, further research is needed on model construction to analyze the effectiveness of more complex deep learning methods.

Funding information: The author states no funding involved.

Author contributions: Conceptualization – Wenjuan Ke; methodology – Wenjuan Ke; formal analysis – Wenjuan Ke; investigation – Wenjuan Ke; resources – Wenjuan Ke; writing-original draft preparation – Wenjuan Ke; writing-review and editing – Wenjuan Ke; visualization – Wenjuan Ke; The author has read and agreed to the published version of the manuscript.

Conflict of interest: No conflict of interest was declared by the author.

Data availability statement: The datasets analysed during the current study are available from the corresponding author on reasonable request.

References

[1] Jat DS, Limbo A, Singh C. Speech-based automation system for the patient in orthopedic trauma ward – ScienceDirect. Smart Biosens Med Care. 2020;201–14.

[2] Berjon P, Nag A, Dev S. Analysis of French phonetic idiosyncrasies for accent recognition. Soft Comput Lett. 2021;3:1–7.

[3] Lazaro JB, Po MCP, Ramones LM, Tolidanes PML. Real-time speech recognition engine for accent correction using hidden markov model. AIP Conference Proceedings, (Bandung, Indonesia); 2018, July 27–28. p. 1–6.

[4] Barkana BD, Patel A. Analysis of vowel production in Mandarin/Hindi/American-accented English for accent recognition systems. Appl Acoust. 2020;162:1–13.

[5] Xiao, B, Kang SC. Development of an image data set of construction machines for deep learning object detection. J Comput Civ Eng. 2021;35:1–18.

[6] Pang K. A decision-making method for self-driving based on deep reinforcement learning. J Phys: Conf Ser. 2020;1576:1–8.

[7] Nahar KMO, Almomani A, Shatnawi N, Alauthman M. A robust model for translating arabic sign language into spoken arabic using deep learning. Intell Autom Soft Comput. 2023;37:2037–57.

[8] Jiang, YQ, Xiong, JH, Li, HY, Yang XH, Yu WT, Gao M, et al. Using smartphone and deep learning technology to help diagnose skin cancer. Br J Dermatol. 2020;182:e95.

[9] Khanam F, Munmun, FA, Ritu NA, Saha AK, Mridha MF. Text to speech synthesis: a systematic review, deep learning based architecture and future research direction. J Adv Inf Technol. 2022;13:398–412.

[10] Zhang W, Zhai M, Huang Z, Li W, Cao Y. Towards end-to-end speech recognition for Chinese mandarin using SE-MCNN-CTC. J Appl Acoust. 2020;39:223–30.

[11] Manohar K, Logashanmugam E. Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm. Knowl Syst. 2022;246:1–22.

[12] Kumar LA, Renuka DK, Rose SL, Shunmuga Priya MC, Wartana IM. Deep learning based assistive technology on audio visual speech recognition for hearing impaired. Int J Cognit Comput Eng. 2022;3:24–30.

[13] Seki H, Yamamoto K, Akiba T, Nakagawa S. Discriminative learning of filterbank layer within deep neural network based speech recognition for speaker adaptation. IEICE Trans Inf Syst. 2019;102:364–74.

[14] Gan Z, Hou M, Hou H, Yang H. Savitzky-Golay filtering and improved energy entropy for speech endpoint detection under low SNR. J Phys: Conf Ser. 2020;1617:1–9.

[15] Syiem B, Dutta SK, Binong J, Singh LJ. Comparison of Khasi speech representations with different spectral features and hidden Markov states. J Electron Sci Technol. 2021;19:155–62.

[16] Heriyanto H, Wahyuningrum T, Fitriana GF. Classification of Javanese script hanacara voice using Mel frequency cepstral coefficient MFCC and selection of dominant weight features. J Infotel. 2021;13:84–93.

[17] Huang Z, Kurotori T, Pini R, Benson SM, Zahasky C. Three-dimensional permeability inversion using convolutional neural networks and positron emission tomography. Water Resour Res. 2022;58:1–21.

[18] Pally RJ, Samadi S. Application of image processing and convolutional neural networks for flood image classification and semantic segmentation. Environ Model Softw. 2022;148:1–15.

[19] Yevnin Y, Chorev S, Dukan I, Toledo Y. Short-term wave forecasts using gated recurrent unit model. Ocean Eng. 2023;268:1–8.

[20] Shobana J, Murali M. An improved self attention mechanism based on optimized BERT-BiLSTM model for accurate polarity prediction. Comput J. 2023;66:1279–94.

[21] Maesa A, Garzia F, Scarpiniti M, Cusani R. Text independent automatic speaker recognition system using mel-frequency cepstrum coefficient and Gaussian mixture models. J Inf Secur. 2012;3:335–40.