Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **10**, Number **4**, November **2025**

Deep Learning-based English Speech Recognition and Pronunciation Quality Evaluation

Yuan-Yuan Li*

School of General Education
Chongqing Jianzhu College
Chongqing 400072, P. R. China
melinda5513@163.com

Shan-Shan Xu

Shandong Huayu University of Technology
Dezhou 253034, P. R. China
Assumption University
Bangkok 10700, Thailand
370739869@qq.com

*Corresponding author: Yuan-Yuan Li
Received September 9, 2024, revised March 24, 2025, accepted September 10, 2025.

Abstract. *Traditional English speech recognition and pronunciation quality evaluation systems perform poorly when dealing with non-native learners' pronunciation and complex noise environments, limiting the effectiveness of their application in English teaching. With the development of deep learning technologies, there is an increasing demand for systems that can simultaneously improve speech recognition accuracy and pronunciation evaluation accuracy. In order to solve these two problems, this paper proposes a multi-task joint learning architecture based on improved deep recurrent neural networks. Firstly, by introducing the attention mechanism and residual connectivity, the model's ability to adapt to complex speech environments is enhanced, and the accuracy and robustness of speech recognition are improved. Second, a multilevel pronunciation assessment framework was designed to combine overall scoring and specific error diagnosis to improve the accuracy of assessing the pronunciation of non-native learners. Finally, the cooptimisation of speech recognition and pronunciation quality evaluation tasks was achieved through a multi-task joint learning architecture. The experimental results show that the word error rates of our model on the LibriSpeech dataset are 3.95% and 9.67% for the "clean" and "other" test sets, respectively. On the L2-ARCTIC dataset, our model has an average absolute error of 0.41, a Pearson correlation coefficient of 0.91, and a diagnostic accuracy of 84.7% in the pronunciation quality evaluation task. The multi-task learning architecture further reduces the word error rate for speech recognition to 3.87%, while reducing the mean absolute error for pronunciation quality evaluation to 0.41.*
Keywords: English speech recognition; pronunciation quality evaluation; deep recurrent neural networks; multi-task learning; non-native learners

1. **Introduction.** With the acceleration of globalisation, the importance of English as an international common language is becoming more and more prominent. In this context, the development of English speech recognition technology and pronunciation quality evaluation systems is of great significance for improving the effectiveness of English teaching [1] and promoting cross-cultural communication [2]. However, existing technologies

still face many challenges in dealing with complex speech environments [3] and evaluating non-native learners' pronunciation [4, 5]. English speech recognition techniques have progressed significantly over the last few decades, from early systems based on Hidden Markov Models (HMM) [6] to current deep learning models. These techniques are capable of achieving high recognition accuracy under ideal conditions, but performance tends to degrade dramatically when faced with background noise, accent variations and large vocabulary tasks. In particular, recognition of English speech for non-native learners [7] is more difficult due to differences in accent and pronunciation habits. Meanwhile, pronunciation quality evaluation systems [8, 9] play an increasingly important role in English language teaching. These systems can provide learners with immediate feedback and help them improve their pronunciation. However, existing evaluation systems are often difficult to accurately assess the pronunciation of non-native learners, and are particularly deficient in capturing subtle pronunciation errors and providing targeted suggestions for improvement. In addition, the two tasks of speech recognition [10, 11] and pronunciation quality evaluation have long been considered as separate issues, researched and developed separately. However, these two tasks are actually closely related in that they both require in-depth analysis and understanding of speech signals. Combining these two tasks is expected to improve overall performance by sharing knowledge and features. In this context, this research aims to address the above challenges through deep learning techniques, in particular improved Deep Recurrent Neural Networks (DRNN) and multi-task learning architectures. Our goal is to develop an integrated system capable of accurately recognising English speech in complex environments and providing accurate pronunciation quality evaluation for non-native learners.

1.1. **Related work.** Existing methods in the field of English speech recognition and pronunciation quality assessment can be analysed from the following three aspects. Firstly, in English speech recognition, traditional methods mainly rely on acoustic models and language models. Gales [12] proposed a speech recognition framework based on HMM, which realises speech-to-text conversion by building acoustic models and language models. This method lays the foundation of statistical speech recognition, but the recognition accuracy still needs to be improved in noisy environments. Ding and Hsu [13] developed a speech recognition algorithm based on dynamic temporal regularisation (DTW) to identify isolated words through template matching. This method performs well in small-vocabulary tasks, but is difficult to be extended to large-vocabulary continuous speech recognition. Secondly, in terms of pronunciation quality assessment, the current mainstream methods are mainly based on feature extraction and scoring models. Franco et al. [14] proposed an automatic pronunciation scoring system based on HMM, which evaluates pronunciation quality by comparing the acoustic characteristics of learners' pronunciation with those of standard pronunciation. This method can give an overall score, but it is difficult to provide a specific diagnosis of mispronunciation. Neumeyer et al. [15] developed a pronunciation network-based scoring method to assess learners' pronunciation by constructing a pronunciation network that contains both correct pronunciations and typical mispronunciations. This method can identify specific mispronunciations, but requires a large amount of expert labelled data. Novotný et al. [16] proposed an automatic scoring method based on duration and spectral features to assess pronunciation fluency by extracting multiple acoustic features. This method performs well in assessing fluency, but still needs to be improved in assessing pronunciation accuracy.

Finally, machine learning techniques show great potential in the field of speech recognition and pronunciation quality evaluation. Graves and Jaitly [17] proposed an end-to-end speech recognition model based on DRNN, which avoids the separation of acoustic and

language models in traditional methods by directly learning the mapping relationship between acoustic features and text. This method achieved a breakthrough in large vocabulary continuous speech recognition task, but the robustness in complex environments still needs to be improved. Wang [18] developed a deep convolutional neural network (DCNN)-based pronunciation quality assessment model to evaluate pronunciation quality by learning acoustic spectrogram features. The study shows that this method outperforms traditional methods in terms of accuracy and consistency, but pronunciation assessment for non-native speakers is still a challenge.

1.2. **Motivation and contribution.** Existing English speech recognition methods mainly rely on traditional techniques such as Hidden Markov Models and Dynamic Time Regulation, which perform poorly in noisy environments and large vocabulary tasks. Meanwhile, current pronunciation quality evaluation methods are usually based on feature extraction and scoring models, which are difficult to provide accurate pronunciation error diagnosis and still face challenges in assessing non-native learners. Furthermore, although machine learning techniques show great potential in these two areas, existing methods tend to treat speech recognition and pronunciation quality evaluation as independent tasks, ignoring the intrinsic connection between them. To address the above issues, this paper proposes a deep learning- based multi-task joint optimisation model to simultaneously improve English speech recognition and pronunciation quality evaluation. The main innovations and contributions of this work include: (1) Aiming at the limitations of traditional speech recognition methods in complex environments, this paper designs an improved DRNN model, which improves the robustness and accuracy of the model in noisy environments and large-vocabulary tasks by introducing the attention mechanism and residual connectivity. This improvement effectively solves the problem of poor performance of end-to-end models in complex environments. (2) To address the challenges of error diagnosis and assessment of non-native learners in pronunciation quality assessment, this paper proposes a multilevel assessment framework based on DCNN, which is able to give both overall scores and specific pronunciation error diagnoses. This method builds on the work of [17] and further improves the accuracy of assessing pronunciation for non-native learners. In addition, in order to take full advantage of the correlation between speech recognition and pronunciation quality evaluation tasks, this paper designs a multi-task joint learning architecture, which achieves co-optimisation of the two tasks by sharing the underlying feature representations and the task-specific upper-layer network, improving the overall performance and reducing the computational complexity.

2. Improved DRNN model.

2.1. **Introduction of attention mechanism.** In order to solve the performance degradation problem of traditional recurrent neural networks when processing long sequences of speech signals, this paper introduces an attention mechanism into the improved DRNN model. This mechanism can effectively capture the key information in the speech signal and improve the robustness and accuracy of the model in noisy environments and large vocabulary tasks.

We define the input speech sequence as $X = (x_1, x_2, \ldots, x_T)$, where T is the length of the sequence and x_t denotes the feature vector at time step t. Conventional recurrent neural networks often struggle to effectively exploit long distance dependencies when dealing with such sequences [19]. For this reason, we design an attention layer for computing the importance weights of each time step.

The attention weight α_t is calculated as follows:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)}. \tag{1}$$

where e_t is a scoring function that measures the relevance of the current hidden state h_t to the input sequence:

$$e_t = v^T \tanh\big(W_h h_t + W_x x_t + b\big), \tag{2}$$

where W_h denotes the weight matrix used to transform the hidden states; W_x denotes the weight matrix used to transform the input features; v is a learnable parameter matrix used to compress the intermediate representations into scalar scores [20]; b denotes a bias term that increases the flexibility of the model.

The design of the scoring function e_t is a central component of the attention mechanism, which aims to quantify the correlation between the current hidden state and each element of the input sequence. The functional structure of Equation (2) employs the structure of a single-layer feed-forward neural network using the hyperbolic tangent (tanh) as the activation function.

The hidden state h_t of the current time step contains the model's memory of past sequences. The input feature x_t of the current time step, represents the current speech information.

The scoring function design idea consists of three main points: (1) Feature fusion: fusion of past and current information is achieved by linearly combining the hidden state and the current input via $W_h h_t + W_x x_t$. (2) Nonlinear transformation: the tanh activation function is used to introduce nonlinearity and enhance the expressiveness of the model, while restricting the value domain to [-1, 1], which facilitates gradient propagation. (3) Dimensional compression: mapping high-dimensional features to scalars via v^T to get the final relevance score.

The designed scoring function is adaptive. By learning the weight matrix, the model can adaptively adjust the attention to different features. Combining h_t and x_t allows the scoring function to take into account both historical information and current inputs, thus enabling context-awareness. In addition, this design allows the model to learn complex non-linear relationships for highly non-linear data such as speech signals. With this design, the scoring function e_t is able to effectively quantify the importance of each element of the sequence to the current prediction, thus allowing the model to dynamically focus on key parts of the input sequence, improving the model's ability to process long sequences of speech signals.

By introducing the attention mechanism, we can obtain a context vector c_t which is a weighted sum of the input sequences:

$$c_t = \sum_{j=1}^{T} \alpha_j x_j. \tag{3}$$

This context vector c_t contains information about the entire input sequence, with special emphasis on the part related to the current time step. We combine c_t with the original hidden state h_t to obtain a new hidden state representation:

$$\tilde{h}_t = \tanh\big(W_c[c_t; h_t] + b_c\big), \tag{4}$$

where W_c is the weight matrix, b_c is the bias term, and $[\,;\,]$ denotes the vector splicing operation. By introducing this attentional mechanism, our model is better able to process long sequences of speech signals, especially in recognising key phonemes and words in noisy environments. This improvement directly addresses the underperformance of

traditional methods in complex environments and improves the model's performance in large vocabulary tasks.

2.2. **Design of residual connectivity.** To further enhance the performance of the model in complex speech environments, we introduced residual connectivity in the improved DRNN [21, 22]. This design aims to solve the problem of gradient vanishing in deep network training, as well as to enhance the model's ability to capture features at different scales. In traditional deep recurrent neural networks, information decays with increasing depth of the network. To overcome this problem, we add residual connections between neighbouring layers. Assuming that the output of the l-th layer is H_l, the output of the $l+1$-th layer can be expressed as:

$$H_{l+1} = F(H_l, W_l) + H_l \tag{5}$$

where $F(H_l, W_l)$ denotes the nonlinear transformation from the l-th layer to the $l+1$-th layer, and W_l is the weight parameter of the layer.

In order to adapt the properties of RNN, we improve the residual connection. Define the hidden state of the l-th layer at time step t as $h_t^{(l)}$, then the improved residual connection can be expressed as:

$$h_t^{(l+1)} = F\big(h_t^{(l)},\, h_{t-1}^{(l+1)},\, W_l\big) + h_t^{(l)}. \tag{6}$$

where $F(h_t^{(l)}, h_{t-1}^{(l+1)}, W_l)$ is a nonlinear transformation that takes into account the current input and the output of the previous time step.

To further enhance the expressive power of the model, we introduce cross-layer residual connections. Define the cross-layer connection function $G(h_t^{(l-k)}, W_g)$, where k denotes the number of layers spanned and W_g is the connection weight. Then the hidden state update method containing cross-layer residual connections is:

$$h_t^{(l+1)} = F\big(h_t^{(l)},\, h_{t-1}^{(l+1)},\, W_l\big) + h_t^{(l)} + G\big(h_t^{(l-k)}, W_g\big). \tag{7}$$

This design allows information to be passed directly between different layers, which helps the model to capture speech features at different scales. The introduction of residual connectivity brings the following advantages: firstly, it alleviates the gradient vanishing problem, making the training of deeper networks more stable. Second, the residual path provides the network with a shortcut for identity mapping, allowing the model to more easily preserve low-level features, which is particularly important for detail preservation in speech recognition. Finally, cross-layer residual connectivity enhances the model's ability to fuse multi-scale features, which facilitates the capture of long and short-term dependencies in speech signals.

This improvement directly responds to the innovation of this paper, which is to improve the robustness and accuracy of the model in noisy environments and large vocabulary tasks by introducing residual connections. Combined with the attention mechanism introduced in the previous paper, the design of residual connectivity further enhances the model's ability to deal with complex speech environments, and lays a solid foundation for achieving high- performance English speech recognition and subsequent pronunciation quality evaluation.

2.3. **Model architecture details.** The overall architecture of the improved DRNN model, which incorporates the attention mechanism and residual connectivity, is designed to improve the performance of English speech recognition, especially in noisy environments and large vocabulary tasks.

This architecture improves robustness to noisy environments by introducing an attention mechanism at each layer that enables the DRNN model to adaptively focus on

important parts of the input sequence. The use of residual connectivity not only solves the gradient problem in deep networks, but also enhances the model's ability to capture multi-scale features, which is particularly important for handling large vocabulary tasks. The overall architecture of the improved DRNN model is shown in Figure 1.

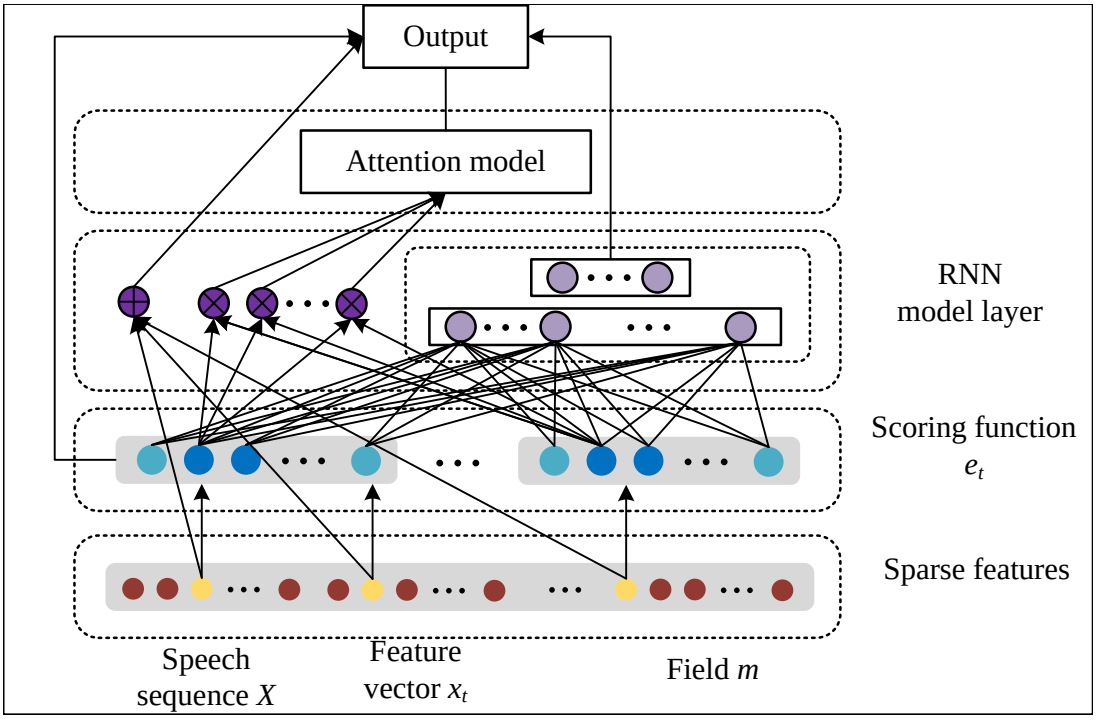

Figure 1. Overall architecture of the improved DRNN model

Our model consists of a multilayer bi-directional long short-term memory network (BiLSTM). Suppose the input speech sequence is $X = (x_1, x_2, \ldots, x_T)$, and the total number of layers of the model is L. For the l-th layer ($1 \le l \le L$), we define the forward and backward hidden states as $h_t^{(l,f)}$ and $h_t^{(l,b)}$, respectively. First, for each layer, we compute the output of the BiLSTM:

$$h_t^{(l,f)} = \mathrm{LSTM}^{(f)}\big(x_t,\, h_{t-1}^{(l,f)}\big), \tag{8}$$

$$h_t^{(l,b)} = \mathrm{LSTM}^{(b)}\big(x_t,\, h_{t+1}^{(l,b)}\big), \tag{9}$$

where $\mathrm{LSTM}^{(f)}$ and $\mathrm{LSTM}^{(b)}$ denote the forward and backward LSTM cells, respectively.

Next, we introduce the attention mechanism. Define the attention weight $\alpha_t^{(l)}$ of the l-th layer as:

$$\alpha^{(l)} = \mathrm{softmax}\Big(v_l^{\top} \tanh\big(W_l^{(h)} [\, h_t^{(l,f)}; h_t^{(l,b)} \,] + W_l^{(x)} x_t + b_l\big)\Big), \tag{10}$$

where v_l, $W_l^{(h)}$, $W_l^{(x)}$, and b_l are the learnable parameters of the layer.

The attention-weighted context vector $c_t^{(l)}$ is computed as follows:

$$c_t^{(l)} = \sum_{j=1}^{T} \alpha_j^{(l)} [\, h_j^{(l,f)}; h_j^{(l,b)} \,]. \tag{11}$$

We then combine the output of the attention mechanism with the original hidden state and apply the residual connection:

$$\tilde{h}_t^{(l)} = \tanh\big(W_l^{(c)} [\, c_t^{(l)}; h_t^{(l,f)}; h_t^{(l,b)} \,] + b_l^{(c)}\big), \tag{12}$$

$$h_t^{(l)} = \tilde{h}_t^{(l)} + [\, h_t^{(l,f)}; h_t^{(l,b)} \,] + G\big(h_t^{(l-k)}, W_l^{(g)}\big), \tag{13}$$

where $W_l^{(c)}$ and $b_l^{(c)}$ are learnable parameters; $G(\cdot, W_l^{(g)})$ is the cross-layer residual connectivity function; and k is the number of layers spanned.

Finally, we use a fully connected layer to map the output of the last layer to a vocabulary-sized space:

$$y_t = \mathrm{softmax}\big(W^{(o)} h_t^{(L)} + b^{(o)}\big), \tag{14}$$

where $W^{(o)}$ and $b^{(o)}$ are the parameters of the output layer; y_t is the predicted vocabulary distribution.

3. **DCNN-based multilevel pronunciation assessment framework.**

3.1. **Overall scoring module.**

In this section, we describe in detail the holistic scoring module in the DCNN-based multilevel pronunciation assessment framework. This module is designed to provide a holistic quality assessment of non-native learners' English pronunciation, which directly responds to the innovation point of this paper, which is to improve the accuracy of the assessment of non-native learners' pronunciation. The overall scoring module adopts a DCNN structure to take full advantage of the time-frequency characteristics of the speech signal. The input is the Mel Frequency Cepstrum Coefficient (MFCC) feature [**?**] of the speech signal, denoted as $M \in \mathbb{R}^{T \times F}$, where T is the number of time steps and F is the frequency dimension.

Firstly, we extract local features of speech using multilayer convolutional layers:

$$C_l = \mathrm{ReLU}\big(W_l * C_{l-1} + b_l\big), \tag{15}$$

where C_l denotes the convolutional output of the l-th layer; W_l and b_l are the convolutional kernel and bias, respectively; $*$ denotes the convolutional operation; ReLU is the activation function. Initial input $C_0 = M$.

To capture speech features at different scales, we use a multi-scale convolutional kernel. Define K different sizes of convolution kernels to get K feature maps:

$$F_k = \mathrm{MaxPool}\big(C_L^{(k)}\big), \quad k = 1, 2, \ldots, K, \tag{16}$$

where $C_L^{(k)}$ is the last layer of convolutional output obtained using the k-th convolutional kernel and MaxPool is the maximum pooling operation.

Next, we introduce an attention mechanism to weight the features at different scales:

$$\alpha_k = \frac{\exp\big(v_a^\top \tanh(W_a F_k + b_a)\big)}{\sum_{j=1}^{K} \exp\big(v_a^\top \tanh(W_a F_j + b_a)\big)}, \tag{17}$$

$$F = \sum_{k=1}^{K} \alpha_k F_k, \tag{18}$$

where W_a, b_a and v_a are parameters of the attention layer; F is the weighted feature representation.

Finally, we use a fully connected layer and a Sigmoid function to map the features to a range of scores from 0 to 100:

$$S = 100 \cdot \mathrm{Sigmoid}(W_s F + b_s), \tag{19}$$

where W_s and b_s are the parameters of the fully connected layer; S is the final overall score.

In order to improve the adaptability of the model to the pronunciation of non-native learners, we introduce an accent adaptive layer. Define the accent feature vector $A \in \mathbb{R}^d$,

where d is the dimension of the accent feature. We fuse the accent features with the speech features:

$$F' = \mathrm{ReLU}\big(W_f[\,F; A\,] + b_f\big), \tag{20}$$

where W_f and b_f are parameters of the fusion layer. The final score is computed using F' instead of F. This design not only captures the multi-scale features of the speech signal, but also highlights the importance of key features through the attention mechanism. The introduction of the accent adaptive layer allows the model to better adapt to non-native learners with different accents, thus improving the accuracy and generalisation of the assessment.

3.2. **Specific pronunciation error diagnosis module.** On top of the overall scoring, we designed the specific pronunciation error diagnosis module to provide precise feedback on pronunciation errors in order to give an overall score and specific pronunciation error diagnosis. This module uses a sequence-to-sequence model based on the attention mechanism. Let the input speech feature sequence be $X = (x_1, x_2, \ldots, x_T)$, and the standard phoneme sequence is $Y = (y_1, y_2, \ldots, y_N)$. Our goal is to generate a sequence of error markers $E = (e_1, e_2, \ldots, e_N)$, where e_i denotes the type of articulatory error for the i-th phoneme.

First, we encode the input sequence using a bidirectional LSTM:

$$h_t = [\,h_t^{(f)}; h_t^{(b)}\,], \quad t = 1, 2, \ldots, T, \tag{21}$$

where $h_t^{(f)}$ and $h_t^{(b)}$ are the hidden states of the forward and backward LSTM, respectively.

In the decoding phase, we use the attention mechanism to generate error markers. For the i-th phoneme, the attention weight is calculated as follows:

$$\alpha_{i,t} = \frac{\exp\big(s_i^\top W_a h_t\big)}{\sum_{j=1}^{T} \exp\big(s_i^\top W_a h_j\big)}, \tag{22}$$

where s_i is the hidden state of the decoder at step i and W_a is the learnable parameter matrix.

The context vector c_i is obtained by weighted summation:

$$c_i = \sum_{t=1}^{T} \alpha_{i,t} h_t. \tag{23}$$

The predicted probability distribution of mislabelling is:

$$p(e_i \,|\, e_{<i}, Y, X) = \mathrm{softmax}\big(W_e[\,s_i; c_i; y_i\,] + b_e\big), \tag{24}$$

where W_e and b_e are output layer parameters; y_i is the embedded representation of the current phoneme.

To improve the accuracy of the diagnosis, we introduce a phoneme similarity matrix $S \in \mathbb{R}^{P \times P}$, where P is the size of the phoneme set. This matrix is used to model the confusion relationship between different phonemes. We integrate this information into the error prediction:

$$\tilde{p}(e_i \,|\, e_{<i}, Y, X) = \mathrm{softmax}\big(W_e[\,s_i; c_i; y_i; S y_i\,] + b_e\big). \tag{25}$$

This design not only accurately identifies specific mispronunciations, but also takes into account similarities between different phonemes, thus providing more precise and targeted diagnostic results.

3.3. **Non-native speaker learner adaptation strategies.** In order to further improve the accuracy of assessing non-native learners' pronunciation, we designed a set of non-native learner adaptation strategies. Firstly, we introduce a learner feature vector $L \in \mathbb{R}^d$, which is used to represent the learner's language background, accent features and other information. This vector is obtained by a pre-trained learner classification model.

We integrate the learner feature vectors into the overall scoring and error diagnosis modules described earlier. For the overall scoring module, the modified feature fusion step is:

$$F'' = \mathrm{ReLU}\big(W_f[\,F'; L\,] + b_f\big). \tag{26}$$

For the error diagnosis module, we introduce learner features at each step of the decoder:

$$s_i = \mathrm{LSTM}\big([\,e_{i-1}; y_i; L\,],\ s_{i-1}\big), \tag{27}$$

where e_{i-1} is the embedded representation of the erroneous tokens predicted in the previous step.

In addition, we design a dynamic threshold adjustment mechanism that dynamically adjusts the threshold of error determination according to the learner's linguistic background. Define the threshold function:

$$\theta(L) = \theta_0 + W_\theta \tanh(V_\theta L + b_\theta), \tag{28}$$

where θ_0 is the base threshold; W_θ, V_θ and b_θ are learnable parameters.

The final rule for error determination is:

$$e_i = \begin{cases} \text{Incorrect.}, & \text{if } \max\big(\tilde{p}(e_i \,|\, e_{<i}, Y, X)\big) < \theta(L), \\ \text{Correct.}, & \text{otherwise.} \end{cases} \tag{29}$$

This non-native learner adaptation strategy allows our assessment system to be personalised to the characteristics of different learners, thus improving the accuracy of the assessment of non-native learners' pronunciation.

4. **Multi-task joint learning architecture.** In order to take full advantage of the intrinsic connection between speech recognition and pronunciation quality evaluation tasks, we propose an innovative multi-task joint learning architecture. This architecture enables the co-optimisation of the two tasks, improving the overall performance and reducing the computational complexity.

4.1. **Shared underlying feature extraction network.** Let the input speech signal be $X \in \mathbb{R}^{T \times F}$, where T is the time step and F is the frequency dimension. The shared feature extraction network can be represented as:

$$H = f_{\text{shared}}(X; \theta_{\text{shared}}), \tag{30}$$

where f_{shared} denotes the operation of the shared network, θ_{shared} is the network parameter, $H \in \mathbb{R}^{T' \times D}$ is the extracted feature, T' is the number of time steps after downsampling, and D is the feature dimension. The shared network contains multilayer convolution, batch normalisation and pooling operations, where the operation at the l-th layer can be represented as:

$$H_l = \mathrm{Pool}\big(\mathrm{ReLU}(\mathrm{BN}(W_l * H_{l-1} + b_l))\big), \tag{31}$$

where W_l and b_l are the convolution kernel and bias of the l-th layer; BN denotes batch normalisation; ReLU is the activation function; and Pool is the pooling operation.

4.2. **Task-specific upper layer network design.** Based on the shared features, we designed task-specific upper layer networks for speech recognition and pronunciation quality evaluation tasks, respectively. The upper layer network f_{asr} for the speech recognition task uses a modified DRNN structure as described in Section 2:

$$Y_{\text{asr}} = f_{\text{asr}}(H; \theta_{\text{asr}}), \tag{32}$$

where Y_{asr} is the predicted text sequence and θ_{asr} is the network parameters.

The upper network of the pronunciation quality evaluation task f_{pq} consists of two sub-modules, overall scoring and error diagnosis, as described in Section 3:

$$Y_{\text{pq}} = f_{\text{pq}}(H; \theta_{\text{pq}}), \tag{33}$$

where Y_{pq} contains the overall score and error diagnosis results, and θ_{pq} is a network parameter.

4.3. **Joint optimisation strategy.** In order to achieve co-optimisation of the two tasks, we design a joint loss function:

$$\mathcal{L} = \lambda_{\text{asr}}\mathcal{L}_{\text{asr}} + \lambda_{\text{pq}}\mathcal{L}_{\text{pq}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{34}$$

where $\mathcal{L}_{\text{asr}}$ and $\mathcal{L}_{\text{pq}}$ are the loss functions for the speech recognition and pronunciation quality evaluation tasks, respectively; $\mathcal{L}_{\text{reg}}$ is the regularisation term; $\lambda_{\text{asr}}, \lambda_{\text{pq}}$ and λ_{reg} are the weighting coefficients to balance the different loss terms.

The loss function for the speech recognition task uses Connectionist Temporal Classification (CTC) loss:

$$\mathcal{L}_{\text{asr}} = -\log p\big(Y^*_{\text{asr}} \,|\, X\big), \tag{35}$$

where Y^*_{asr} is a sequence of real texts.

The loss function for the pronunciation quality evaluation task consists of a mean square error loss for overall scoring and a cross-entropy loss for error diagnosis:

$$\mathcal{L}_{\text{pq}} = \mathcal{L}_{\text{score}} + \mathcal{L}_{\text{diag}}, \tag{36}$$

$$\mathcal{L}_{\text{score}} = \frac{1}{N}\sum_{i=1}^{N}\big(S_i - S_i^*\big)^2, \tag{37}$$

$$\mathcal{L}_{\text{diag}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}^* \log(p_{ij}), \tag{38}$$

where S_i and S_i^* are the predicted and true overall scores, respectively; y_{ij}^* is the true mislabelling; p_{ij} is the predicted probability of error; N is the number of samples; and M is the number of error categories.

To facilitate knowledge transfer between two tasks, we introduce an inter-task attention mechanism. Defining the inter-task attention matrix $A \in \mathbb{R}^{D\times D}$, the feature transformation can be expressed as:

$$H_{\text{asr}} = H + \alpha\, AH, \qquad H_{\text{pq}} = H + (1-\alpha)A^{\top}H, \tag{39}$$

where α is the learnable equilibrium parameter; H_{asr} and H_{pq} are the post-transformation features for speech recognition and pronunciation quality evaluation tasks, respectively.

With this multi-task joint learning architecture, we achieve the co-optimisation of speech recognition and pronunciation quality evaluation tasks. Sharing the underlying feature extraction network not only reduces computational complexity, but also facilitates knowledge migration between the two tasks. The task-specific upper layer network design ensures that the specific needs of each task are met, while the joint optimisation strategy balances the learning objectives of the different tasks. This design provides

a unified framework for achieving efficient and accurate English speech recognition and pronunciation quality evaluation.

5. **Experiments and analysis of results.**

5.1. **Datasets and experimental setup.** In order to comprehensively evaluate our proposed deep learning-based English speech recognition and pronunciation quality evaluation system, we selected several representative datasets and designed a series of experiments. These experiments aim to verify the performance of the proposed model in noisy environments and large-vocabulary tasks, as well as the accuracy of evaluating the pronunciation of non-native learners.

We used the following dataset:
(1) LibriSpeech dataset [25]: contains 1000 hours of English speech data for evaluating a large vocabulary continuous speech recognition task. We use the "clean" and "other" subsets to test the performance of the model under different acoustic conditions.
(2) TIMIT dataset [26]: contains English speech data from 630 speakers and is used to evaluate the recognition accuracy at the phoneme level.
(3) L2-ARCTIC dataset [27]: contains speech data from non-native English learners from multiple language backgrounds to assess pronunciation quality evaluation tasks.
(4) Self-constructed noise environment dataset: we added different types and intensities of noise to LibriSpeech and constructed a noise environment test set containing 5000 samples.

The experimental setup is as follows:
(1) Speech feature extraction: we use 40-dimensional Meier frequency cepstrum coefficients (MFCC) as input features with a frame length of 25ms and a frame shift of 10ms.
(2) Model parameters: the shared underlying feature extraction network contains 5 convolutional layers using 256 3×3 convolutional kernels per layer. The upper network for the speech recognition task uses 3 layers of bi-directional LSTM with 512 hidden units per layer. The upper layer network for the pronunciation quality evaluation task uses 2 layers of bi-directional LSTM with 256 hidden units per layer.
(3) Training setup: we use the Adam optimiser with an initial learning rate set to 0.001 and a batch size of 64. A learning rate decay strategy is used during training, where the learning rate is halved when the validation set performance does not improve for 3 consecutive epochs. We train on 4 NVIDIA Tesla V100 GPUs for a total of 50 epochs.
(4) Evaluation metrics: for speech recognition tasks, we use word error rate (WER) and phoneme error rate (PER) as evaluation metrics. For the pronunciation quality evaluation task, we use Mean Absolute Error (MAE), Pearson Correlation Coefficient (PCC), and Diagnostic Accuracy (DA) as evaluation metrics.

5.2. **Speech recognition performance evaluation.** First, we evaluate the performance of our model on the LibriSpeech dataset for large-vocabulary continuous speech recognition. Figure 2 shows the WER results of our model compared to several baseline models on the "clean" and "other" test sets.

Our model achieves the lowest word error rates on both the "clean" and "other" test sets. Particularly noteworthy is that our model reduces the error rate on test-clean and test-other by 1.15% and 4.53%, respectively, compared to the end-to-end DRNN model. This significant improvement proves that our proposed improved DRNN model addresses the problem of poor performance of end-to-end models in complex environments. Next, we evaluate the robustness of the model on a self-constructed noisy environment dataset. We compared the WER at different SNRs and the results are shown in Figure 3.

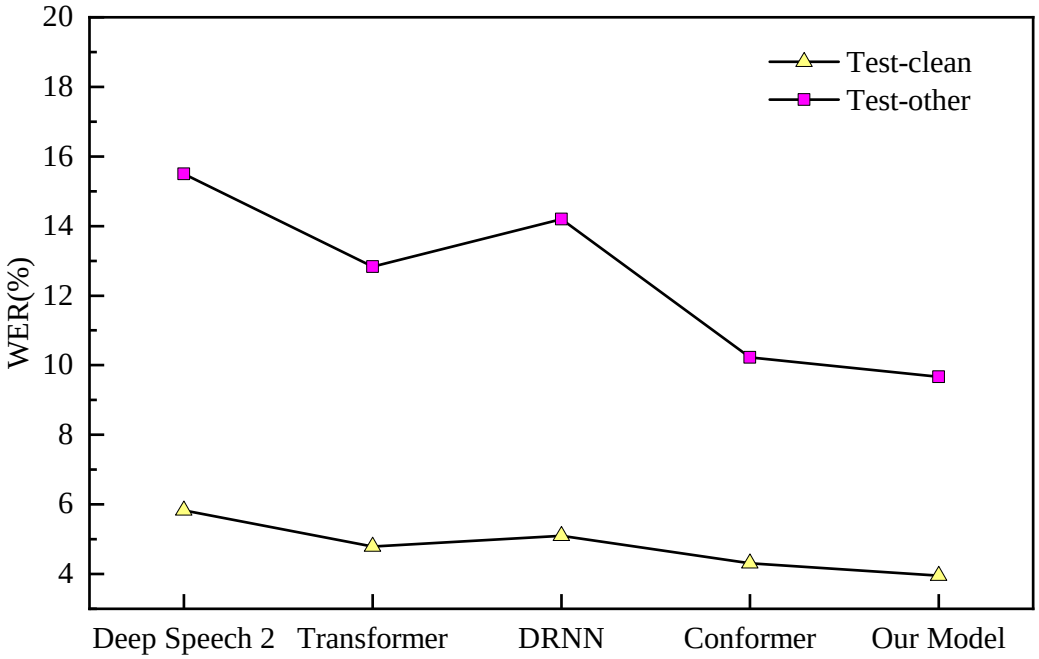

Figure 2. WER on LibriSpeech dataset (%).

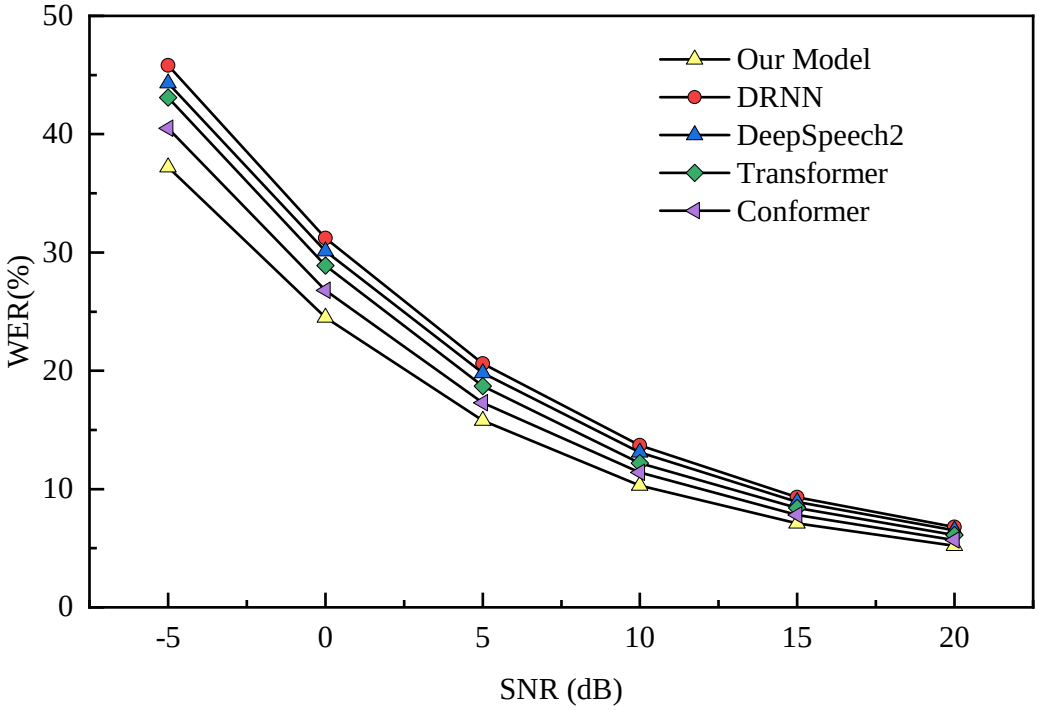

Figure 3. WER at different SNR (%).

Our model still maintains the lowest word error rate at all noise levels. The Conformer model outperforms the other baseline models (DRNN, DeepSpeech2, and Transformer), but still slightly underperforms our proposed model, which shows the IMPROVEMENT of our model with respect to DRNN. Under low SNR conditions (such as 0 dB and -5 dB), our model still has a significant advantage over Conformer, reducing the error rate by 2.3% and 3.3%, respectively.

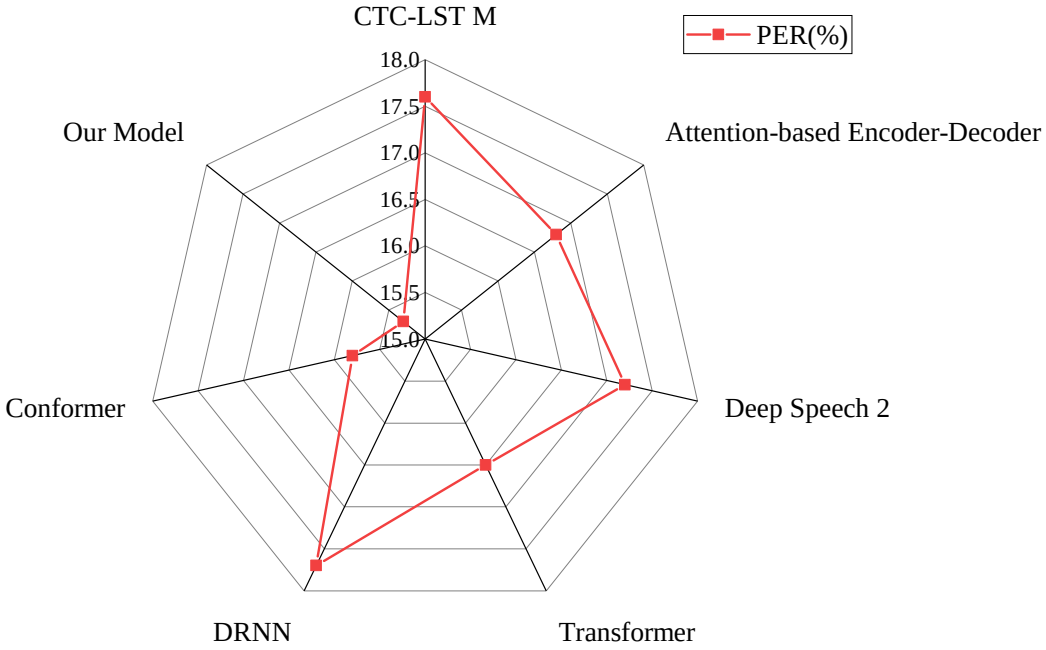

Figure 4. PER on the TIMIT dataset (%).

To further validate the phoneme level recognition ability of the models, we evaluated them on the TIMIT dataset. The comparison results of PER are presented in Figure 4. CTC-LSTM and Attention-based Encoder-Decoder are the commonly used baseline models on the TIMIT dataset. The results show that our model achieves the lowest error rate of 15.3% on the phoneme-level recognition task. It reduces the error rate by 2.4% compared to DRNN, and there is also a 0.5% IMPROVEMENT compared to the latest Conformer model.

5.3. **Analysis of the effectiveness of pronunciation quality assessment.** We will evaluate the performance of our proposed model on a pronunciation quality evaluation task, with a special focus on its accuracy in assessing the pronunciation of non-native learners. We mainly use the L2-ARCTIC dataset for our experiments. We used Mean Absolute Error (MAE), Pearson Correlation Coefficient (PCC) and Diagnostic Accuracy (DA) as the evaluation metrics. MAE reflects the average difference between the model's predicted scores and the human ratings, PCC measures the correlation between the predicted scores and the human ratings, and DA denotes the model's accuracy in identifying the specific articulation errors. Table 1 demonstrates the performance of the different models on the L2-ARCTIC dataset.

Table 1. Pronunciation Quality Evaluation Results on the L2-ARCTIC Dataset

Model	MAE	PCC	DA (%)
DCNN	0.58	0.82	76.3
DNN-based	0.53	0.85	78.9
LSTM-based	0.49	0.87	80.5
Our Model	0.41	0.91	84.7

As can be seen in Table 1, our model achieves the best results on all three metrics. Compared to DCNN, our model reduced MAE by 0.17, improved PCC by 0.09, and

improved DA by 8.4%. These results fully demonstrate the superiority of our model in assessing the pronunciation quality of non-native learners.

5.4. **Analysis of the effect of joint multi-task learning.** In order to evaluate the effectiveness of multi-task learning, we compared the performance of the single-task model and the multi-task model on both tasks. Table 2 shows the comparison results on the LibriSpeech and L2-ARCTIC datasets.

Table 2. Single Task vs. Multi-Task Model Performance Comparison

Model	Libri Speech WER (%)	L2-ARCTIC MAE	L2-ARCTIC PCC
Single-ASR	3.95	–	–
Single-PQ	–	0.46	0.89
Multi-task	3.87	0.41	0.91

As can be seen in Table 2, the multi-task model achieved better performance on both tasks. On the speech recognition task, the WER of the multi-task model was reduced by 0.08% compared to the single-task model. On the pronunciation quality evaluation task, the multi-task model achieved a 0.05 reduction in MAE and a 0.02 improvement in PCC. This improvement can be attributed to the knowledge migration between the two tasks. These results directly validate the innovation proposed in this paper, i.e., the co-optimisation of speech recognition and pronunciation quality evaluation tasks by means of a multi-task joint learning architecture. Our architecture not only improves the performance of the two tasks, but also reduces the overall computational complexity, since the two tasks share most of the network parameters. In summary, our proposed multi-task joint learning architecture successfully achieves the co-optimisation of speech recognition and pronunciation quality evaluation tasks, and enhances the generalisation ability and computational efficiency of the model while improving the performance. This provides a powerful and efficient solution for the development of English speech teaching systems.

6. **Conclusion.** In this paper, a multi-task joint learning architecture based on improved deep recurrent neural networks is proposed for English speech recognition and pronunciation quality evaluation, which effectively solves the limitations of traditional methods in dealing with non-native learners' pronunciation and complex noise environments. By introducing the attention mechanism and residual connectivity, the model is able to capture the key features of speech signals more accurately, which significantly improves the accuracy and robustness of speech recognition. In addition, the multilevel pronunciation evaluation framework combines overall scoring and specific error diagnosis to further enhance the ability to assess the pronunciation of non-native learners and ensure the accuracy of the evaluation results. By conducting experiments on LibriSpeech and L2-ARCTIC datasets, the following conclusions can be drawn:
(1) The improved deep recurrent neural network model can significantly improve the performance of speech recognition in complex environments, especially for non-native accents.
(2) The multilevel pronunciation assessment framework provides a more accurate and detailed evaluation of pronunciation quality than traditional methods, especially for diagnosing pronunciation errors in non-native learners.
(3) The multi-task joint learning architecture achieves efficient use of computational resources while improving the overall performance, which is the optimal strategy recommended in this paper.

The experiments in this paper are mainly based on the LibriSpeech and L2-ARCTIC datasets, which, although representative, may not fully cover all possible speech variants and learner backgrounds. In addition, although our model has made significant progress in processing non-native learner pronunciation, there may be room for improvement in the model's performance for extreme accents or very rare pronunciation errors.

REFERENCES

[1] H. W. Kam, "English language teaching in East Asia today: An overview," Asia Pacific Journal of Education, vol. 22, no. 2, pp. 1-2, 2002. pp. 1-22, 2002.

[2] Y. A. Wu, "English language teaching in China: Trends and challenges," Tesol Quarterly, pp. 191-194, 2001.

[3] R. Akbari, F. Behzadpoor, and B. Dadvand, "Development of English language teaching reflection inventory," System, vol. . 38, no. 2, pp. 211-227, 2010.

[4] A. Kirkpatrick, "Teaching English across cultures. What do English language teachers need to know to know how to teach English," English Australia Journal, vol. 23, no. 2, pp. 20-32, 2007. " English Australia Journal, vol. 23, no. 2, pp. 20-36, 2007.

[5] T. Pica, "Tradition and transition in English language teaching methodology," System, vol. 28, no. 1, pp. 1-18, 2000.

[6] E. Trentin and M. Gori, "A survey of hybrid ANN/HMM models for automatic speech recognition," Neurocomputing, vol. 37, no . 1-4, pp. 91-126, 2001.

[7] M. H. Long, "Native speaker/non-native speaker conversation and the negotiation of comprehensible input1," Applied Linguistics, vol. 4, no. 2, pp. 126-141, 1983.

[8] Y. Xu, "English speech recognition and evaluation of pronunciation quality using deep learning," Mobile Information Systems, vol. 2022, no. 1, 7186375, 2022.

[9] Y. Hong and H. Nam, "Evaluating score reliability of automatic English pronunciation assessment system for education," Studies in Foreign Language Education, vol. 35, no. 1, pp. 91-104, 2021.

[10] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," Multimedia Tools and Applications, vol. 80, pp. 9411-9457, 2021.

[11] H. Aldarmaki, A. Ullah, S. Ram, and N. Zaki, "Unsupervised automatic speech recognition: a review," Speech Communication, vol. 139, pp. 76-91, 2022.

[12] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Computer Speech & Language, vol. 12, no. 2, pp. 75-98, 1998.

[13] I.-J. Ding and Y.-M. Hsu, "An HMM-Like Dynamic Time Warping Scheme for Automatic Speech Recognition," Mathematical Problems in Engineering, vol. 2014, no. 1, 898729, 2014.

[14] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in 1997 IEEE International Conference on Acoustics, Speech, And Signal Processing. ieee, 1997, vol. 2, pp. 1471-1474.

[15] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," Speech communication, vol. 30, no. 2-3, pp. 83-93, 2000.

[16] M. Novotný, J. Rusz, R. Čmejla, and E. Růžička, "Automatic evaluation of articulatory disorders in Parkinson's disease ," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 9, pp. 1366-1378, 2014.

[17] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in International Conference on Machine Learning. pMLR, 2014, pp. 1764-1772.

[18] L. Wang, "English Speech Recognition and Pronunciation Quality Evaluation Model Based on Neural Network," Scientific Programming, vol. 2022, no. 1, 2249722, 2022.

[19] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," Neurocomputing, vol. 452, pp. 48-62,. 2021.

[20] M.-H. Guo et al. "Attention mechanisms in computer vision: a survey," Computational Visual Media, vol. 8, no. 3, pp. 331 -368, 2022.

[21] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," Computers, Materials & Continua, vol. 79, no. 1, pp. 19-46, 2024.

[22] Y. Ma, Y. Peng, and T.-Y. Wu, "Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning," Journal of Intelligent & Fuzzy Systems, vol. 43, no. 2, pp. 2121-2133, 2022.

[23] O. C. Ai, M. Hariharan, S. Yaacob, and L. S. Chee, "Classification of speech dysfluencies with MFCC and LPCC features," Expert Systems with Applications, vol. 39, no. 2, pp. 2157-2165, 2012.
[24] X. Wang, S. Yin, H. Li, J. Wang, and L. Teng, "A network intrusion detection method based on deep multi-scale convolutional neural network," International Journal of Wireless Information Networks, vol. 27, pp. 503-517, 2020.
[25] O. H. Anidjar, R. Marbel, and R. Yozevitch, "Harnessing the power of Wav2Vec2 and CNNs for Robust Speaker Identification on the VoxCeleb and LibriSpeech Datasets," Expert Systems with Applications, vol. 255, 124671, 2024.
[26] D. Byrd, "Preliminary results on speaker-dependent variation in the TIMIT database," The Journal of the Acoustical Society of America, vol. 92, no. 1, pp. 593-596, 1992.
[27] Z. Zhang, Y. Wang, and J. Yang, "Text-conditioned transformer for automatic pronunciation error detection," Speech Communication, vol. 130, pp. 55-63, 2021.