Journal of Network Intelligence ©2025 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **10**, Number **2**, May **2025**

A Quality Evaluation Model for College English Teaching Based on Tone Detection and Deep Learning

Chun-Yan Peng*

Zhejiang Yuexiu University, Shaoxing 312000, P. R. China
Graduate School, Lyceum of the Philippines University, Intramuros, Manila 1002, Philippines
`20242031@zyufl.edu.cn`

Xiang-Zhen Zhou

Zhejiang Yuexiu University, Shaoxing 312000, P. R. China
Faculty of Information Science and Technology, National University of Malaysia, Selangor 43600, Malaysia
`619543699@qq.com`

*Corresponding author: Chun-Yan Peng

Received August 28, 2024; revised January 27, 2025; accepted March 29, 2025.

Abstract. *With the continuous advancement of educational informatization, the evaluation of the quality of college English teaching has become a key link in improving teaching effectiveness and students' language abilities. Traditional methods for evaluating teaching quality often rely on subjective evaluation, which makes it difficult to comprehensively and objectively reflect the actual teaching situation. Therefore, this article proposes a university English teaching quality evaluation model based on tone detection and deep learning. This method first uses a convolutional network module to learn the spatial features abstracted from the Mel frequency cepstral coefficient feature map of the input speech. Then, the convolutional layer is added to the residual network for superposition to improve the feature extraction capability of the model. Finally, considering the strong correlation of speech signal context, a Transformer module is introduced into the model to capture the temporal information and global features of the tested speech, thereby improving the overall performance of the model. The experimental results show that the proposed model can effectively capture subtle changes in students' speech with high accuracy and has good application prospects. This study provides a new intelligent method for evaluating the quality of college English teaching, which helps to further enhance the scientific and effective nature of English teaching.*

Keywords: tone detection; deep learning; English language teaching

1. **Introduction.** Driven by globalization, English has become the main language for international communication, making it particularly important to improve the quality of university English teaching. However, the existing teaching quality evaluation system has many shortcomings, mainly reflected in excessive reliance on subjective evaluation, which makes it difficult to comprehensively and objectively reflect teaching effectiveness [1,2]. This not only limits teachers' accurate assessment of students' language abilities, but also hinders the improvement and innovation of teaching methods. With the continuous advancement of educational informatization and the rapid development of artificial intelligence technology, intelligent evaluation methods based on voice data have gradually become a research hotspot. In the field of speech evaluation, tone, as one of the important

features of speech, can reflect the speaker's emotions, tone, and language fluency. Tone detection not only plays an important role in pronunciation assessment of language learning, but also provides new perspectives and methods for objective evaluation of teaching quality. However, traditional tone detection methods often rely on rule-based algorithms, which exhibit limitations when dealing with complex speech features [3,4].

To address this issue, deep learning technology has been introduced into the field of speech processing, achieving automatic extraction and analysis of complex speech features through advanced models such as Convolutional Neural Networks (CNN) and Transformers. This article proposes an innovative model for evaluating the quality of college English teaching by combining tone detection and deep learning techniques. Specifically, this article first uses convolutional neural networks to extract Mel frequency cepstral coefficients (MFCC) feature maps from student speech, in order to capture spatial features in the speech. Next, this article integrates convolutional layers into a residual network (ResNet) and further improves the accuracy and efficiency of feature extraction by superimposing residual modules. Finally, considering the temporal and context dependent nature of speech signals, this paper introduces the Transformer module to capture the global features and temporal information of the tested speech, thereby comprehensively improving the performance and accuracy of the model. The experimental results show that the evaluation model proposed in this paper has significant advantages in processing complex speech data, and can effectively capture subtle changes in speech, with accurate and consistent evaluation results. Compared to traditional subjective evaluation methods, this model not only improves the objectivity and scientificity of evaluation, but also provides an intelligent solution for improving the quality of English teaching. Through this model, teachers can have a more comprehensive understanding of students' language learning situation and adjust teaching strategies accordingly, further optimizing teaching effectiveness.

The following chapters will provide a detailed introduction to the model construction, experimental process, and result analysis of this study. Firstly, a detailed introduction will be given to the relevant theoretical techniques involved in this article, mainly introducing two English assessment methods, including referenced assessment and non referenced assessment. Then, research was conducted on the English assessment method based on residual networks and Transformers. Finally, the optimal composition structure and parameters of the model were obtained through experiments, and compared with other network models to demonstrate the effectiveness of the proposed method.

1.1. **Related work.** The current English evaluation is mainly based on the quality of pronunciation. Pronunciation quality evaluation refers to the use of machines. The device evaluates the pronunciation quality of the test speech, or language experts subjectively evaluate the test speech based on corresponding evaluation indicators. Subjective scoring is mainly based on aspects such as pronunciation content, prosodic characteristics, and perceptual characteristics, which to some extent reflect the subjective perception of the human ear on the quality of the tested speech pronunciation.

Objective pronunciation quality evaluation is mainly divided into pronunciation quality scores with reference and pronunciation quality scores without reference. The reference based pronunciation quality evaluation methods mainly use speech recognition technology or the standard pronunciation model trained by Hidden Markov Model (HMM) to automatically match and discriminate the speech of the detected person. Reference based pronunciation quality evaluation refers to providing the examinee with a standard reference speech that has been judged by certain criteria during pronunciation evaluation, allowing the examinee to learn and repeat pronunciation. The current mainstream pronunciation evaluation method usually uses speech recognition technology to model pronunciation, and evaluates the pronunciation of the detected person by identifying the number of incorrect pronunciations in their speech. The earliest method based on speech recognition used confidence calculation to segment and decode the detected speech and corresponding text using forced alignment. Using the intermediate results obtained from decoding, the likelihood ratio of the HMM model was calculated to evaluate the pronunciation of the tested speech. Witt and Young [5] did not use the HMM posterior probability method, but took a different approach by using phonemes, aligning and segmenting phonemes, and finding syllables with pronunciation errors to evaluate the deviation between the detector's pronunciation and the standard speech, in order to achieve the goal of pronunciation quality evaluation. In addition, Witt also elaborated on the Goodness of Pronunciation (GOP) algorithm. The basic idea of the GOP algorithm is to use known text to forcibly align the tested speech with its corresponding text (Force Alignment, FA), and then compare the obtained likelihood score with the likelihood score obtained without corresponding text. This likelihood ratio (LR) is used as the standard for evaluating the quality of pronunciation. The GOP algorithm is implemented using speech recognition technology and requires a large amount of data to train the acoustic model, typically recordings of native speakers. Since the GOP algorithm was proposed, it has been widely used in various computer-aided pronunciation training systems (CAPT).

Although many scholars have conducted in-depth research on it, most existing pronunciation evaluation algorithms have been improved based on the GOP algorithm. Witt et al. [6] continued to conduct in-depth research on GOP, which is essentially the ratio of likelihood scores. When Witt used GOP for pronunciation quality scoring, different decoding methods were used for the divisor of the ratio, with the numerator using a forced alignment network and the denominator using a cyclic network of phonemes. However, the final result was not ideal and the accuracy of the scoring decreased. In addition, there are researchers who use maximum classification and minimum phoneme error methods to evaluate pronunciation for phonemes. In addition to conducting research on improving the GOP algorithm, some researchers have applied the CAPT system to the learning of The Second Language. Reference [7] has expanded the pronunciation dictionary. Many scholars have also conducted pronunciation evaluation research on Chinese.

Since the beginning of the 21st century, with the rapid development of machine learning technology in the field of speech, both machine learning and deep learning have gradually been used in pronunciation evaluation systems. The GOP method has been widely used in CAPT systems since 2000. But as the amount of data increases, research has found that Support Vector Machine (SVM) performs better than GOP. Wei et al. [8] used an SVM classifier to train a vector machine for each phoneme, achieving higher accuracy than GOP and better false pronunciation detection results. Fu et al. [9] also used the DNN-HMM model and evaluated speech using Reference free Error Rate (RER). Piotrowska et al. [10] used Convolutional Neural Networks (CNN) and LSTM to find methods for detecting English aspirated and other allophones, in order to achieve automatic pronunciation evaluation. In recent years, Transformer has been successfully applied in the field of natural language processing and has received widespread attention in the speech industry. Reference [11] is a study on using Transformer for automatic error detection in speech.

1.2. **Contribution.** At present, the English pronunciation evaluation methods based on deep learning mainly evaluate the tested speech based on standard pronunciation acoustic models. The current mainstream pronunciation model is based on speech recognition technology, which evaluates the tested speech by identifying the number of incorrect pronunciations in the detected speech. However, this method requires extremely high accuracy of the recognition model, requires a large dataset for training the model, and only detects incorrect pronunciation in sentences, ignoring tone issues. Therefore, this method cannot comprehensively evaluate the English proficiency of the test taker. To address this issue, this paper investigates a new network model to comprehensively evaluate the English pronunciation of test takers and improve the correlation between output results and artificial labels. The main contributions of this article are as follows:

(1) This article proposes a university English teaching quality evaluation model based on tone detection and deep learning: the model integrates modern speech processing technology and deep learning algorithms, which can effectively evaluate the quality of university English teaching.

(2) This article innovatively combines convolutional networks and residual networks: using dilated convolution modules to learn abstract spatial features of spectrograms, and expanding the receptive field, so that the output of convolution can map a larger range of information, and overlaying it in the residual network improves the feature extraction ability of the model, thereby enhancing the accuracy of speech signal analysis.

(3) This article introduces the Transformer module: considering the contextual relevance of speech signals, the model incorporates the Transformer module to capture the temporal information and global features of speech, further improving the overall performance of the model.

(4) The experiment verified the effectiveness of the model: it was shown through experiments that the proposed model can accurately capture subtle changes in students' speech with high accuracy, demonstrating good application prospects.

2. Theoretical analysis.

2.1. English pronunciation evaluation with reference.

Reference based English speech evaluation refers to comparing the recorded speech of the test subject with a standard reference speech of the same text. Common methods include acoustic features, prosodic features, perceptual features, as well as segment duration and fluency of long sentences. The method based on acoustic features mainly evaluates the test speech by comparing the similarity of acoustic features between standard speech and the test speech. The resonance peak is essentially the resonant frequency on the acoustic cavity, which to some extent reflects the important physical characteristics of the vocal tract. According to its common evaluation method, HMM logarithmic likelihood scoring is used to evaluate the test speech based on standard pronunciation as a reference [12, 13]. If a phoneme is used as the basic unit for evaluating the tested speech, and t_i represents the time when the i-th phoneme starts to be pronounced, then the score can be expressed as:

$$l_i = \sum_{\tau=t_i}^{t_i} \log \left[p(q_\tau \mid q_{\tau-1}) p(o_\tau \mid q_\tau) \right] \tag{1}$$

where, o_τ and q_τ are the observation vectors and HMM states at time, $p(q_\tau \mid q_{\tau-1})$ is the transition probability, and $p(o_\tau \mid q_\tau)$ is the output probability distribution of q_τ. Afterwards, the scores of all phoneme evaluations are summed up to obtain the score of a single word or sentence. Due to the different lengths of words and sentences, they can have a certain impact on the evaluation. Therefore, all phoneme scores are adjusted according to length, that is:

$$G = \frac{\sum_{i=1}^{N} l_i}{\sum_{i=1}^{N} d_i} \tag{2}$$

where, N is the total number of phonemes in the sentence, and $d_i = \tau_{i+1} - \tau_i$ is the duration frame of the i-th element. This processing can result in phonemes with shorter durations being covered by phonemes with longer durations, and some short phonemes may contain important perceptual characteristics. Therefore, evaluation scores are often corrected using local averages, which can be expressed as:

$$\tilde{G} = \frac{1}{N}\sum_{i=1}^{N}\frac{l_i}{d_i} \tag{3}$$

Due to the irregular pronunciation between different learners, this method is greatly influenced by learners, and the accuracy of evaluation and artificial similarity are also low. Compared to HMM logarithmic likelihood, logarithmic posterior probability does not undergo drastic changes with the personal characteristics or vocal tract changes of the detected individual, making it the most commonly used evaluation method currently.

Given the phoneme r_i and its associated i-segment speech frame observation vector o_t, the posterior probability can be expressed as:

$$P(r_i \mid o_t) = \frac{P(o_t \mid r_i)P(r_i)}{\sum_{j=1}^{M} P(o_t \mid r_j)P(r_j)} \tag{4}$$

where, $P(o_t \mid r_i)$ is the probability distribution of phoneme r_i in observation vector o_t, $P(r_j)$ is the prior probability of r_i and M is the number of phonemes in the database that are unrelated to the evaluation text. In the i-th speech, the posterior probabilities of each frame are logarithmically processed and accumulated to obtain the evaluation score of phoneme r_i in the i-th speech, which is:

$$P_i = \sum_{t=\tau_i}^{\tau_{i+1}-1} \log\left[P(r_i \mid o_t)\right] \tag{5}$$

where, τ_i represents the starting time of r_i in speech segment i. The evaluation score for a word or sentence is obtained by taking the logarithmic posterior probability of all phoneme segments, then normalizing them according to phoneme length and taking the average:

$$P = \frac{1}{N}\sum_{i=1}^{N}\frac{p_i}{d_i} \tag{6}$$

In addition to the criterion of resonance peak, the rhythm of speech is also one of the important evaluation criteria. The rhythm of speech refers to the pattern of changes in the duration of speech. From the perspective of acoustic characteristics, it can be further divided into sound intensity, pitch, and length. And the parameters corresponding to them are stress, fundamental frequency, and segment length. The fundamental frequency refers to the oscillatory relaxation vibration of the vocal cords caused by airflow passing through the glottis, resulting in periodic impacts. The methods for extracting the fundamental frequency of speech can be basically divided into time-domain and frequency-domain methods. Since the fundamental frequency does not exist in every frame, it is necessary to perform unvoiced/voiced decision while calculating the fundamental frequency. The time-domain method takes the time-domain waveform of the tested speech as input and finds the minimum positive period of its waveform. Therefore, the signal can be shifted to find the point with the highest overlap with the original signal. There are two methods to measure the coincidence of signals: autocorrelation function and average amplitude difference function.

(1) Auto Correlation Function (ACF) The autocorrelation function multiplies the translated signal with the original signal, and then calculates the sum of a frame of signals. The peak of the correlation function corresponds to the signal shift τ, where W is the number of sampling points within the frame, which can be expressed as:

$$r_t(\tau) = \sum_{i=t}^{t+w-1} x_i x_{i+\tau} \tag{7}$$

(2) Average Magnitude Difference Function (AMDF) The average amplitude difference function is obtained by subtracting the translated signal from the original signal, taking the absolute value of the result, and then summing it up. It can be expressed as:

$$D_t(\tau) = \sum_{i=t}^{t+w-1} |x_i x_{i+\tau}| \tag{8}$$

And integer multiples of the fundamental frequency will have peaks in the spectrum, so the basic principle of the frequency domain method is to find the greatest common divisor of the peak frequency. The general method for obtaining the fundamental frequency is to design a function for each test f to determine the likelihood that f is f_0, which is the Salience Function. The common frequency domain methods are as follows:

(1) Harmonic Summation The harmonic addition method is that the amplitude of a speech signal has a peak at an integer multiple of the fundamental frequency f_0. Therefore, the sum of the integer multiples of f_0 on the amplitude spectrum represents the significance of f_0, which is expressed as follows:

$$H(f) = \sum_{k=1}^{n} |X(kf)| \tag{9}$$

(2) Average Peak to Valley Distance (APVD) The expression for the average peak valley distance is:

$$d_k(f) = \big|X(kf)\big| - \frac{1}{2}\big|X\big((k-\frac{1}{2})f\big)\big| - \frac{1}{2}\big|X\big((k+\frac{1}{2})f\big)\big| \tag{10}$$

$$D_k(f) = \frac{1}{2}\sum_{k=1}^{n} d_k(f) \tag{11}$$

where, $d_k(f)$ represents the significance of the peak at kf relative to the adjacent valley, which is the peak valley distance.

In the process of manual evaluation, in addition to evaluating objective content such as pronunciation and rhythm, due to the unique perceptual characteristics of the human ear, the signals received by the evaluator are not exactly the same as the actual signals [14]. At present, the auditory characteristics of the human ear are limited to research in psychoacoustics and linguistics. The main auditory characteristics include timbre, loudness, pitch, and masking effect. Therefore, perceptual characteristics can also be used as one of the measures for English evaluation. Common perceptual features include Bark spectrum, MFCC spectrum, etc. [15].

2.2. **English pronunciation evaluation without reference.** The evaluation of pronunciation quality without reference is mainly based on a standard pronunciation acoustic model to assess the pronunciation of the examinee. Therefore, the accuracy of the pronunciation model is crucial for the evaluation results. The current mainstream pronunciation models are constructed based on HMM models. In the traditional GMM-HMM model, the probability space of each state component of HMM is described by a Gaussian Mixture Model (GMM) containing numerous components [16]. With the rapid development of deep learning technology, the application scope of speech recognition is becoming increasingly wide. Various neural networks have replaced GMM models and achieved higher accuracy in the field of pronunciation evaluation. The DNN-HMM model is commonly used [17].

3. **English speech evaluation method based on residual network and Transformer.** The network structure diagram of Res Prepared Transformer proposed in this article is shown in Figure 1. The network is mainly composed of dilated convolution modules, Res BN Relu modules, Transformer modules, and fully connected layers. Due to the fact that the tested speech is divided into multiple rating levels based on the English test scoring criteria, and in any rating level, the number and quantity of non-standard English pronunciation problems in each tested speech are not the same, the features of the input MFCC spectrogram are also very abstract. Therefore, a convolution module is needed to learn the abstract spatial features of spectrograms to facilitate model evaluation. Due to the ability of dilated convolutional layers to expand the dimensionality of feature maps and increase the receptive field, the preprocessed data first passes through the dilated convolutional layer.

Next, the data needs to go through N layers of Res BN Relu modules. The introduction of this module is as follows: The feature map passed through the dilated convolutional layer is input into the Res BN Relu layer, which first undergoes 1×1 convolution. Its function is to map low dimensional features to high-dimensional space, expanding the dimensionality of the feature map. Afterwards, the BN layer is used to accelerate the network training speed, and the activation function is used to fit the nonlinear data. Next, the features of the feature map are extracted through a 3×3 convolution. After passing through the BN layer and relu function, the features are compressed using a 1×1 convolution structure to map high-dimensional features onto a low dimensional space. The final output residual mapping is added to the directly mapped portion after 1×1 convolution, and then subjected to pooling and dropout layers to enhance the model's generalization ability.

In addition, in addition to the feature information on the spectrogram, the time series characteristics of the tested speech for each rating level are also complex and diverse. Therefore, a Transformer module was added to the model to capture the temporal information of the tested speech and better learn global features, making the model's judgment of the tested speech more accurate. Finally, after passing through the max pooling layer and fully connected layer, the network output result is obtained.

3.1. **Residual network.** In the process of using neural networks, in order to achieve better training results, the network is often deepened to obtain more hierarchical features. However, as the number of layers increases, ordinary neural networks become difficult to train, and may even experience problems such as gradient explosion, gradient disappearance, and network degradation, leading to a decrease in network performance. The Residual Network (ResNet) has to some extent solved this problem. The structure of the residual block is shown in Figure 2.

Assuming the input of the neural network is x, the output directly mapped is $F(x)$, and the mapping structure is "convolution activation function convolution". x is added to the mapped $F(x)$ and then passed through the activation function to obtain the output of the residual structure. By adding the input to the output, the integrity of the information is effectively preserved, and the loss and damage of information are reduced to a certain

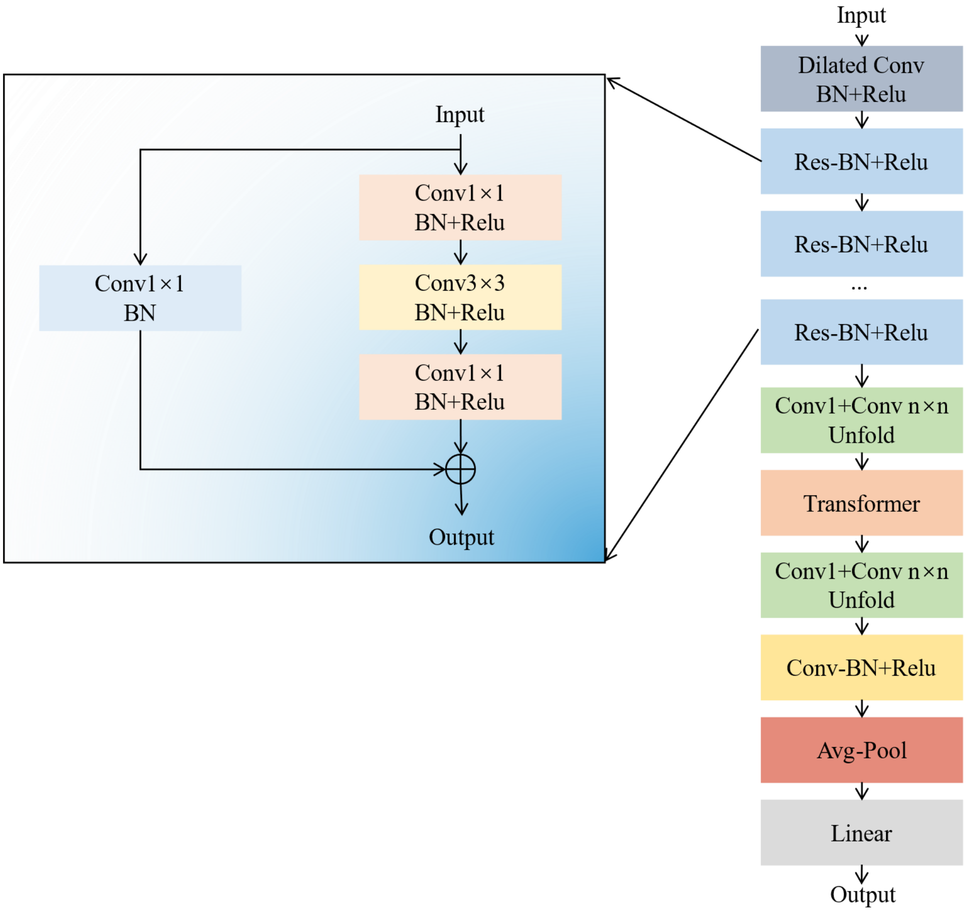

Figure 1. Structure of Res-Dilated-Transformer model

extent [19]. In addition, compared with ordinary networks, residual networks learn $F(x) = H(x) - x$ instead of $H(x)$. Since they only learn the difference between input and output, the learning objectives and difficulty of the network are also reduced.

3.2. **Dilated convolution.** In image processing, pooling or downsampling layers are often used to increase the receptive field of the image. But as the dimensionality of the data decreases and the network parameters decrease, while the receptive field increases, the resolution decreases and some image details are lost. Dilated convolution solves the problem of missing information in pooled images [20].

Dilated convolution is the addition of a dilated rate parameter on the basis of convolution. Dilated convolution is essentially the distance between each convolution kernel during image processing [21]. The kernel size of a regular convolution is 3 × 3 and the hole ratio is 1. The size of the dilated convolution kernel is 3 × 3, and the dilation rate is 2. From this, it can be seen that dilated convolution can expand the receptive field and preserve the details of the image without being lost. The true size of the convolution kernel is shown below:

$$k^* = k + (k - 1)(r - 1) \tag{12}$$

where, k is the original convolution kernel size, k^* is the true convolution kernel size when the hole rate is greater than 1, and r is the hole rate.

3.3. **Transformer network.** The Transformer network was first proposed by Google in 2017 and was mainly used in the field of natural language processing [22]. With the

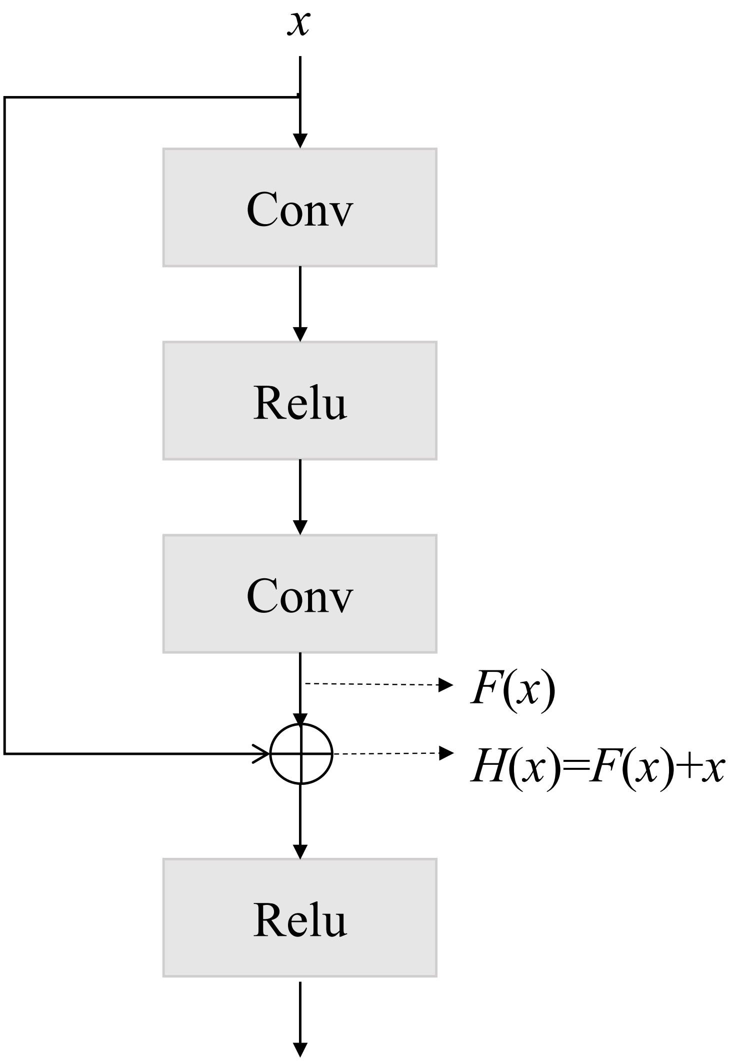

Figure 2. Structure of diagram of residual block

continuous development of deep learning, it has gradually been applied to other fields as well. Its advantage lies in the fact that the method of extracting sequence information is different from traditional RNNs, using attention mechanism for sequence extraction, which can achieve parallel data processing and effectively improve the problem of slow network training in traditional RNNs. The structure of Transformer is shown in Figure 3. Below is a brief introduction to the model.

(1) Input. The input of Transformer is sequence information. Taking natural language processing as an example, after the input embeddings, the position encoding of each word vector needs to be added. It is generated by different sine and cosine functions, and its function is to label each position of the input sequence. The expression is as follows:

$$PE(pos, 2i) = \sin \frac{pos}{1000^{\frac{2i}{d_{model}}}} \tag{13}$$

$$PE(pos, 2i+1) = \cos \frac{pos}{1000^{\frac{2i}{d_{model}}}} \tag{14}$$

where, pos represents the absolute position and d_{model} represents the vector dimension.

(2) Encoder. The encoding part consists of N encoders. The important component of Encoder is the Multi Head Attention mechanism, which calculates correlations. It first generates three weight matrices according to the self attention mechanism, and then

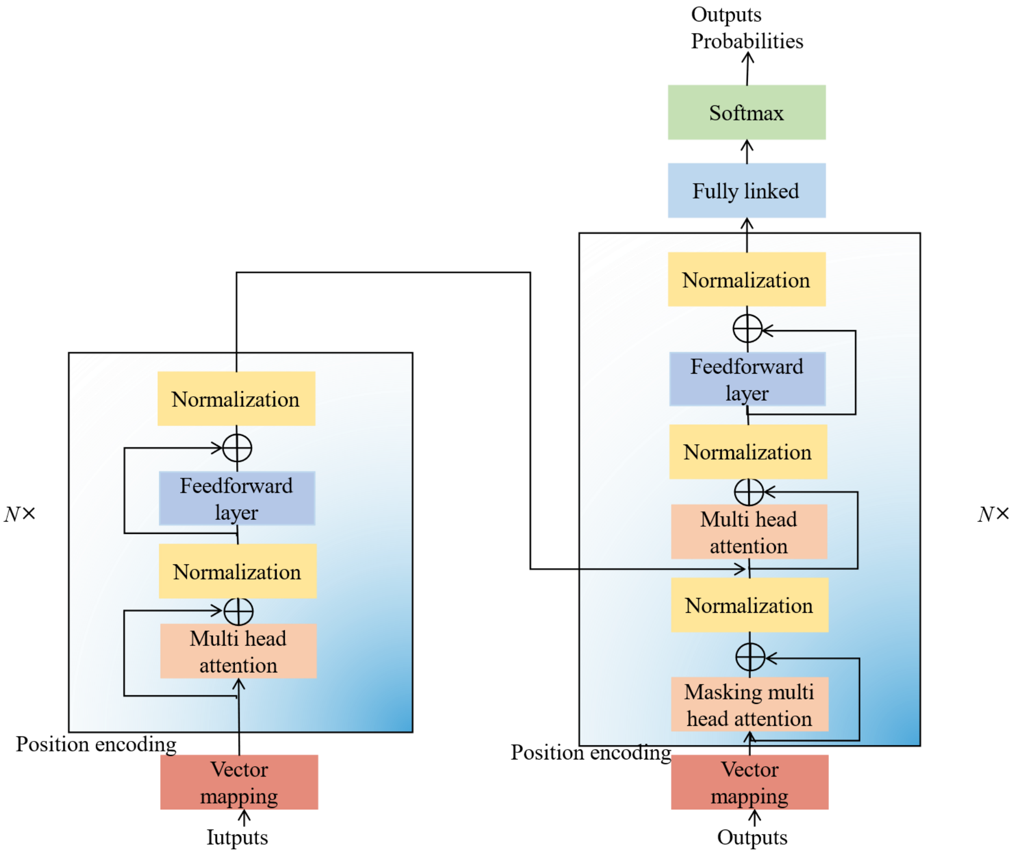

Figure 3. Structure of Transformer

calculates the attention matrix through QK^T. The calculation formula for weight and attention is as follows:

$$Q = X_{embedding}W_Q \tag{15}$$

$$K = X_{embedding}W_K \tag{16}$$

$$V = X_{embedding}W_V \tag{17}$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{18}$$

where $\sqrt{d_k}$ is to transform the Attention matrix into a standard normal distribution, and after softmax normalization, the sum of weights is 1. The normalization module mainly involves residual linking and transforming hidden layers into standard normal distributions to accelerate convergence and prevent network degradation. The function of the feedforward layer is to activate the input through linear mapping.

(3) Decoder The decoder's input consists of the previous output and the encoder's encoding, and its structure is similar to the encoding part. The difference is the use of Masked processing on the first multi head attention, which prevents the decoder from receiving future information during training.

(4) Output The output part of the decoder undergoes linear transformation and then softmax to obtain the probability distribution of the output. Due to the fact that Transformer uses attention mechanism to extract sequence information, it can capture global

contextual information, thus enabling the construction of more distant dependencies on global targets and better capturing global features. But the disadvantage is that it has a large number of parameters and requires a large dataset for training. Therefore, this article uses the lightweight network module MobileViT Block, which combines the advantages of convolutional networks and self attention mechanisms.

The Transformer module in the MobileViT Block module is different from the original Transformer module with the structure shown in Figure 3. The Transformer in MobileViT only uses the Encoder module, and the normalization part of the Transformer is before the multi head attention and MLP modules. MLP consists of layer normalization, fully connected layers, activation functions, and fully connected layers. The original Transformer's ability to obtain local information is not as good as CNN and RNN, and it requires more parameters to fit the global information between patches, which makes the entire network more complex and increases the number of parameters. The MobileViT Block module used in this article combines convolution with the original Transformer, which can better obtain local information, effectively capture global features, and achieve the effect of reducing the number of parameters with better performance.

4. **Experimental results and analysis.**

4.1. **Dataset.** This article aims to evaluate the quality of English teaching for college students. Due to the lack of a dedicated open-source dataset for this article's evaluation, the dataset used in this article was recorded using professional recording equipment according to experimental requirements. The dataset is named DUT-SPEECH, consisting of 15 hours and 25 speakers, including 13 male speakers and 12 female speakers, with a total of over 8000 recorded voices. The recording location is a quiet room without noise, with a signal-to-noise ratio of over 30dB. The duration of a single speech in the dataset is controlled within 7 seconds. Speech with a recording duration of less than 7 seconds is subsequently processed with zero padding to facilitate model learning through neural networks. The sampling frequency of all voice data is 16KHz, and PCM encoding is 16 bits. In the experiment, the dataset was divided into 80% training dataset and 20% testing dataset, with no cross corpus between the two sets. The duration of a single voice in the dataset is controlled within 7 seconds, and voice recordings with a duration of less than 7 seconds are subsequently processed with zero padding. The MFCC dimension extracted from preprocessing is 40, with a frame length of 30ms and a frame shift of 10ms. The window function is a Hamming window. The network optimizer used in the experiment is ADAM algorithm [**?**], the loss function is binary cross entropy loss function, the epochs are set to 100, the minimum training batch is set to 64, and the learning rate is 0.0001.

4.2. **English evaluation method based on Res Compiled Transformer model.** The English evaluation system based on the Res Available Transformer model is shown in Figure 4, which consists of two parts: data preprocessing module and Res Available Transformer model. Among them, the data preprocessing module mainly extracts MFCC features, including pre emphasis, frame windowing, Fourier transform, Mel filtering, and discrete cosine transform. The Res Compiled Transformer model mainly classifies input feature maps according to evaluation criteria.

The speech in the data is divided into four levels, each level containing multiple situations, and the types of input speech feature maps are also relatively complex. Therefore, the model needs to learn the complex features of each rating level. The system first preprocesses the input speech data, extracts the MFCC feature map, reads the corresponding manual evaluation labels of the speech, binds them, and sends them together to the Res

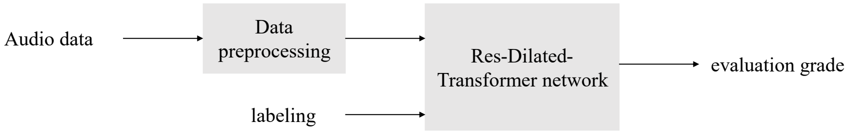

Figure 4. Mandarin assessment system based on Res-Dilated-Transformer network

Compiled Transformer model for evaluation. After a series of processing by the model, the input features are finally classified based on the evaluation of the tested speech grade.

This article uses accuracy as the primary evaluation metric. Represented by Acc, accuracy is the ratio of correctly classified samples to the total number of samples and is widely used in classification tasks. The calculation formula is provided below:

$$Acc = \frac{TP + TN}{N} \tag{19}$$

where TP represents false negative, TN represents true negative.

4.3. **Simulation experiment and result analysis.** This section mainly studies the influence of various hyperparameters of the Res Compiled Transformer baseline network on model performance on the DUT-SPEECH dataset, and compares the proposed model with other reference models through experiments. Accuracy and parameter count are used as evaluation indicators for the model.

In this experiment, Res Compiled Transformer was compared with CNN, RNN, LSTM, VGGNet, ResNet, AlexNet and DenseNet, where CNN and LSTM are divided into medium and large models. The dataset used is DUT-SPEECH, with the same data preprocessing and MFCC parameters, Res Compiled Transformer. There are 1 Dilated Conv module, 4 Res BN Relu modules, and 1 MobileViT Block module, placed after the second Res BN Relu module. The Transformer has 2 multi head attention heads, 3 encoding modules, 32 input channels for the feedforward link layer, and 40 output channels for the feedforward link layer. Under the above experimental parameters, the comparison experiments between the Res Compiled Transformer model and other network structures are shown in Table 1.

Table 1. Comparative experimental results

Model	*Acc*	**Parameter quantity**
CNN-M	0.7465	310
CNN-L	0.7763	890
LSTM-M	0.6134	133
LSTM-L	0.5805	532
CRNN-M	0.7513	371
VGG	0.8083	1092
ResNet	0.7954	684
AlexNet	0.8040	1067
DenseNet	0.7844	1134
Res-Dliated-Transformer	**0.8245**	**600**

From Table 1, it can be seen that the performance of Res Prepared Transformer is superior to other models. Compared with basic models such as CNN and RNN, the accuracy has been significantly improved; Compared with the model in CNN, the accuracy has improved by nearly 8%; Compared to LSTM, there is a maximum improvement of

22%. Compared with the ResNet series of networks, the accuracy of the model is 2.9% higher than that of ResNet. When compared with the VGG series of networks, the performance improvement of the model is not high, with an accuracy increase of up to 0.2%, but the number of parameters has been reduced. Compared with VGG, the number of parameters has been reduced by 331K. Compared with AlexNet and DenseNet, the Res Compiled Transformer model has improved accuracy and reduced parameter count, saving computational resources.

5. **Conclusion.** This article proposes an innovative model for evaluating the quality of college English teaching, aimed at addressing the subjectivity and shortcomings of traditional evaluation methods. This model utilizes convolutional networks to extract speech features and enhances the depth and accuracy of feature extraction through residual networks. In addition, the model also introduces the Transformer module, which can effectively capture temporal information and global features in speech signals, ensuring a more comprehensive and accurate evaluation. The experimental results show that the model performs well in capturing subtle changes in student speech with high accuracy, demonstrating broad application prospects. Overall, this study provides an intelligent solution for evaluating the quality of college English teaching, significantly enhancing the scientific and effective nature of the assessment and playing an important role in promoting English teaching reform.

REFERENCES

[1] 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.

[2] M. Li, "Multidimensional analysis and evaluation of college English teaching quality based on an artificial intelligence model," *Journal of Sensors*, vol. 2022, Art. ID 1314736, 2022.

[3] C.-P. Chu, L.-T. Shen, and S.-H. Hwang, "A new algorithm for tone detection," *AASRI Procedia*, vol. 8, pp. 118–122, 2014.

[4] J.-G. Gander, "A pattern recognition approach to tone detection," *Signal Processing*, vol. 1, no. 1, pp. 65–81, 1979.

[5] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2–3, pp. 95–108, 2000.

[6] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, 2001.

[7] Z. Shufang, "Design of an automatic English pronunciation error correction system based on radio magnetic pronunciation recording devices," *Journal of Sensors*, vol. 2021, Art. ID 5946228, 2021.

[8] S. Wei, G. Hu, Y. Hu, and R.-H. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models," *Speech Communication*, vol. 51, no. 10, pp. 896–905, 2009.

[9] J. Fu, Y. Chiba, T. Nose, and A. Ito, "Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models," *Speech Communication*, vol. 116, pp. 86–97, 2020.

[10] M. Piotrowska, A. Czyżewski, T. Ciszewski, G. Korvel, A. Kurowski, and B. Kostek, "Evaluation of aspiration problems in L2 English pronunciation employing machine learning," *The Journal of the Acoustical Society of America*, vol. 150, no. 1, pp. 120–132, 2021.

[11] Z. Zhang, Y. Wang, and J. Yang, "Text-conditioned transformer for automatic pronunciation error detection," *Speech Communication*, vol. 130, pp. 55–63, 2021.

[12] H. Hamada, S. Miki, and R. Nakatsu, "Automatic evaluation of English pronunciation based on speech recognition techniques," *IEICE Transactions on Information and Systems*, vol. 76, no. 3, pp. 352–359, 1993.

[13] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," *Computer Speech & Language*, vol. 21, no. 1, pp. 219–230, 2007.

[14] D. Mu, W. Sun, G. Xu, and W. Li, "Japanese pronunciation evaluation based on DDNN," *IEEE Access*, vol. 8, pp. 218644–218657, 2020.

[15] C. Molina, N. B. Yoma, J. Wuth, and H. Vivanco, "ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion," *Speech Communication*, vol. 51, no. 6, pp. 485–498, 2009.

[16] 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.

[17] Q. Chen, X. Wang, P. Su, and Y. Yao, "Auto adapted English pronunciation evaluation: a fuzzy integral approach," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 22, no. 1, pp. 153–168, 2008.

[18] H. Alaeddine and M. Jihene, "Deep residual network in network," *Computational Intelligence and Neuroscience*, vol. 2021, Art. ID 6659083, 2021.

[19] Z. Sheng, H. Wang, G. Chen, B. Zhou, and J. Sun, "Convolutional residual network to short-term load forecasting," *Applied Intelligence*, vol. 51, no. 4, pp. 2485–2499, 2021.

[20] S. Tang, J. Xia, L. Fan, X. Lei, W. Xu, and A. Nallanathan, "Dilated convolution based CSI feedback compression for massive MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 10, pp. 11216–11221, 2022.

[21] J. Zhang, C. Lu, J. Wang, L. Wang, and X.-G. Yue, "Concrete cracks detection based on FCN with dilated convolution," *Applied Sciences*, vol. 9, no. 13, Art. ID 2686, 2019.

[22] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.