International Journal for Housing Science and Its Applications
Publish August 4, 2025. Volume 46, Issue 3 Pages 1507-1520

https://doi.org/10.70517/ijhsa463116

Study on the Improvement of Speech Recognition Algorithm Based on Attention Mechanism in English Listening Training under the Background of Smart Education

Li Xia[1,*]
[1] Department of Public Course Teaching, Nanyang Vocational College of Agriculture, Nanyang, Henan, 473000, China
Corresponding authors: (e-mail: xiali@nyca.edu.cn).

Abstract Intelligent education is an important direction of future education, and it is imperative to carry out English listening training teaching in the context of intelligent education. Aiming at the problems existing in traditional speech recognition algorithms, we first preprocess the speech data, and then use the MFCC algorithm to complete the work of English speech feature extraction, take the speech features as the input of the DCNN-CTC model, and after the continuous optimization and training of the attention mechanism, we finally design a speech recognition algorithm based on the ASKCC-DCNN-CTC model, and then validate the algorithm Analysis. On the basis of the speech recognition algorithm based on the DCNN-CTC model, after adding the ASKCC attention mechanism, the UA value is enhanced by 2.96% improvement, while the WA value is also increased by 4.44%, which verifies the improvement of the ASKCC attention mechanism on the speech recognition algorithm.

Index Terms MFCC, DCNN-CTC, ASKCC, speech recognition algorithm, english listening training

I. Introduction

The importance of English has become more and more obvious in today's trend of information and economic globalization, and the use of information technology-assisted teaching has also become a trend in all kinds of schools. The use of multimedia resources to stimulate English classroom teaching, as well as various teaching platforms built in the network environment, is conducive to maximizing the value of teaching resources and stimulating students' interest in learning [1], [2]. At the same time, a new round of teaching and learning reform is also in full swing with the times relying on network resources and modern information means [3].

As an important tool for international communication, the cultivation of English listening ability is particularly important. Traditional English listening teaching methods often have problems such as single classroom form and untimely feedback, which make it difficult to effectively improve students' listening comprehension [4], [5]. The introduction of intelligent speech recognition technology into the English listening classroom can convert students' speech into text in real time and compare it with standard pronunciation, thus realizing sentence-by-sentence listening and feedback [6], [7]. The core advantage of this technology lies in its efficiency and accuracy, which can instantly recognize students' specific problems in listening and pronunciation and provide targeted corrective suggestions [8], [9]. Through intelligent speech recognition, students are not only able to obtain instant feedback and correct their errors in time, but also understand their listening and pronunciation weaknesses through systematic data analysis [10]-[12].

On the one hand, the use of intelligent speech recognition technology can significantly improve students' listening comprehension and pronunciation accuracy.Li, D. et al. developed a set of English pronunciation teaching methods using speech recognition technology, which is conducive to improving students' pronunciation accuracy and listening comprehension by focusing on the changes that occur in their vocabulary and phrases [13]. Wang, Y. constructed an auxiliary training system for English listening and speaking skills based on a binary decision tree, and by inputting students' listening training data into this model, they can be accurately assessed in terms of pronunciation and other aspects, which in turn improves students' English listening and speaking skills [14]. Mirzaei, M. S. et al. explored the help of automatic speech recognition technology in listening training for L2 learners, which can help students to locate difficult speech regions and deepen their listening comprehension of difficult speech segments by identifying errors within the segments [15].

On the other hand, intelligent speech recognition technology can also provide students with personalized learning feedback and suggestions.Liu, J. examined the role of machine learning technology in evaluating the effectiveness of adaptive English listening training, which provided students with an adaptive and personalized learning

experience by building an evaluation model based on the IPSO-BP network, significantly improving their English listening skills [16]. Liu, Y. and Quan, Q. studied the method of English pronunciation error recognition in the process of personalized English learning, and established a Hidden Markov Model of speech recognition to provide students with reliable feedback of pronunciation information, which helps to target the correction of students' pronunciation errors in listening training [17]. Jingning, L. showed that combining mobile sensor networks with speech recognition algorithms to build an English assisted learning system will realize an effective English learning tool by accurately recognizing students' listening practice conversations while providing good learning feedback and interactive experience [18].

Drawing on related information and literature, it is found that there are three kinds of problems in traditional speech recognition algorithms, phoneme features are not standardized, information leakage, and key feature extraction difficulties. Firstly, from the aspect of speech data preprocessing, in the use of MFCC algorithm for feature extraction of English speech data that has been preprocessed, deploy the feature data as an input to the DCNN-CTC model, and after continuous training and optimization of the ASKCC attention mechanism, the speech recognition algorithm improvement based on the attention mechanism is finally realized. The relevant model parameters are set, and under the guidance of evaluation indexes, the verification analysis of speech recognition algorithm based on ASKCC-DCNN-CTC model is completed, which in turn promotes the high-quality development of English listening teaching in the context of smart education.

II. Research on the Improvement of Recognition Algorithm for English Listening Training

II. A. Description of the problem

In recent years, artificial intelligence represented by deep learning has developed rapidly, and a variety of software and functions continue to appear. Many of these deep learning applications are closely related to English teaching. In English listening training, speech recognition has made great progress, and the accuracy rate is constantly improving. In the context of the information age, with the continuous promotion and popularization of robots, a wide variety of speech recognition human-computer interaction systems appear on the market, and there are still problems such as unstandardized phoneme features, information leakage, and difficulties in extracting key features.

II. A. 1) Unstandardized phoneme characteristics

Each person has his or her own unique timbre and multiple ways of expressing speech, for example, in the two contexts of sadness and happiness, the way of pronunciation is obviously different, in order to achieve a higher accuracy of the recognition effect, it is necessary to make the model to adapt to more complex and flexible way of speaking. Therefore, it is necessary for the model to realize the phoneme feature standardization operation in the process of feature extraction.

II. A. 2) Information leakage

The prediction accuracy of the neural network model is largely affected by the quality of the training dataset, and the traditional CNN will have the problem of feature information leakage when performing convolutional computation because it utilizes the data from future frames, which will lead to the model using the data from future frames during training, which in turn will affect the prediction accuracy of the model in the prediction stage.

II. A. 3) Difficulty in key feature extraction

With the improvement of the recognition accuracy requirements for neural network models and the increase in the number of training samples, it is difficult for neural network models to capture key feature information during the training phase. In this paper, we flexibly introduce an effective attention mechanism network for feature fusion to achieve the purpose of feature information splicing in the channel dimension, so that the model can extract more key channel features that play a decisive role in model prediction.

II. B. Speech data preprocessing

To carry out subsequent processing of audio files, the first thing to do is to carry out preprocessing operations on the audio data. The pre-processing process of the speech signal is very important in the speech recognition system. The pre-processing process can ensure the smoothness, uniformity and distortion-free nature of the subsequently processed speech signals, and eliminate the effects of aliasing, high-frequency and high-harmonic distortions of the speech acquisition equipment on the quality of the speech signals, so that the quality of the final speech recognition can be improved. The pre-processing technique of speech signal generally includes the following steps: pre-emphasis, frame splitting and windowing, which are briefly introduced next.

II. B. 1) Speech pre-emphasis

The transfer function of the first-order filter often used for speech signals is equation (1) to accomplish the pre-emphasis operation of speech data [19]. To wit:

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

In Equation (1), α represents the pre-emphasis coefficient of speech data, which has a certain value range, generally take $0.9 < \alpha < 1$, in this paper, the value of α is 0.97. The relationship between the input and output speech signals of the pre-emphasis process is shown in Equation (2):

$$y(t) = x(t) - \alpha x(t-1) \tag{2}$$

where $x(t)$ denotes the original unprocessed speech signal and $y(t)$ denotes the speech signal after the pre-emphasis operation. An audio data in the dataset is selected and the pre-emphasis operation is performed on it, Fig. 1 is the original waveform graph and Fig. 2 is the waveform graph after pre-emphasis.

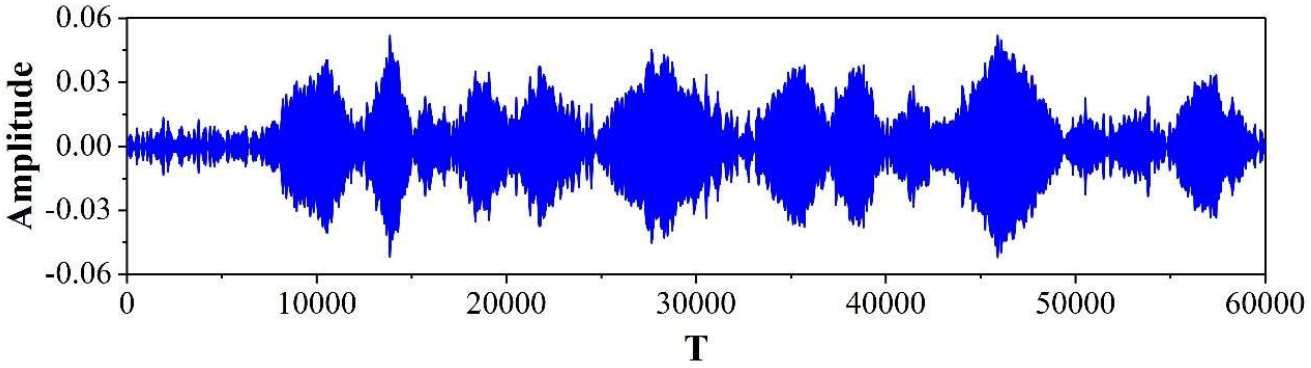


Figure 1: Original waveform

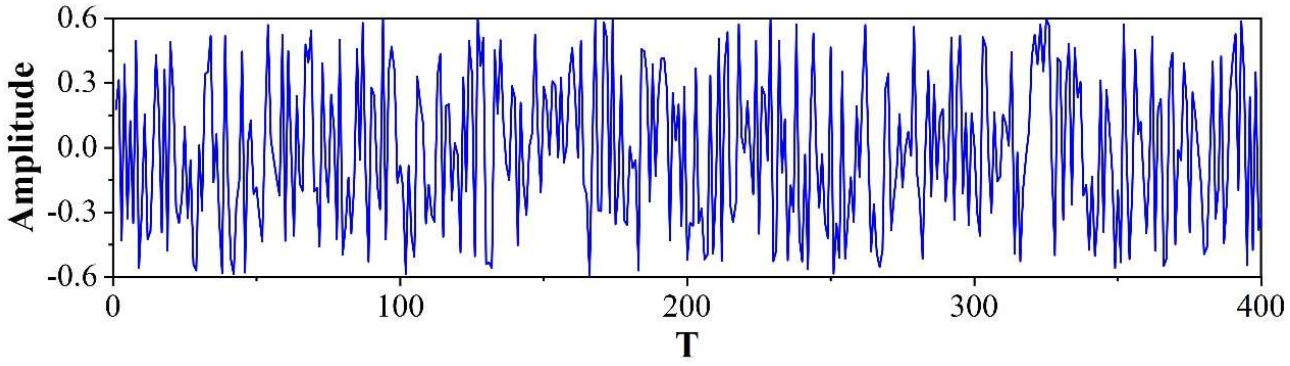


Figure 2: Speech pre-accentuated waveform

II. B. 2) Speech frame processing

Figure 3 gives a schematic diagram of the frame shift and frame length, and Figure 4 shows the subframe waveform. Since the characteristics of the speech signal are continuous and unsteady, even after digitization, its frequency domain will still behave differently at different moments, but in a very small period of time (20-60ms) the audio signal can be considered stable, i.e., the short-time smoothness of the speech signal. According to this principle, in this short period of time can be considered as a quasi-steady state process of the audio signal fragment, the use of sub-frame mode will be each frame length of 20ms or 25ms, so that each frame can be analyzed for the characteristic parameters, and then after the combination of the overall voice signal characteristics of the characteristic values of the characteristic parameters of the composition of the time series.

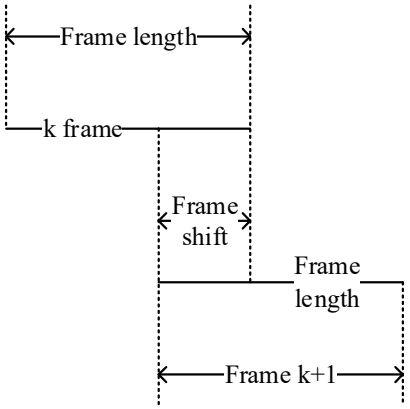


Figure 3: Speech signal framing diagram

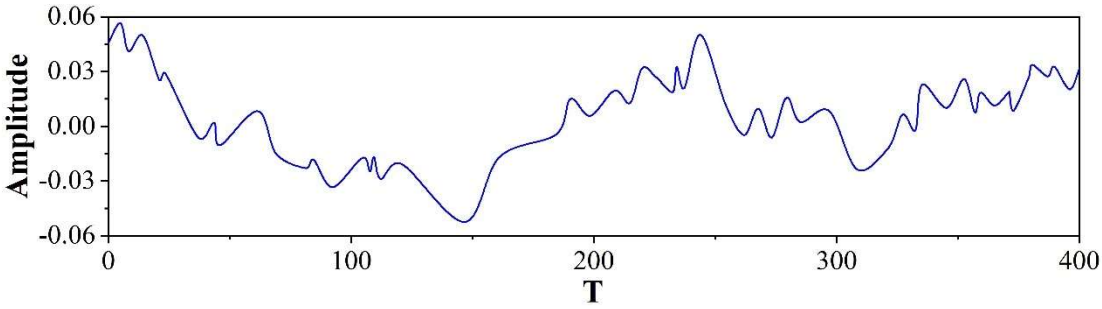


Figure 4: Frame waveform diagram

II. B. 3) Speech windowing

After the sub-frame processing of the speech file, the individual segments after the sub-partitioning of some kind of transformation or with some kind of computation to be processed, that is, its for the addition of window processing, in simple terms, is to divide a section of the speech signal into a one by one small window to be processed, you can get the so-called short-time smooth signal. The function of rectangular window is:

$$W_{reg}(n)=\begin{cases}1, 0\le n\le N-1\\ 0, n\le 0, n\ge N-1\end{cases} \tag{3}$$

where N is the length of the window.

Usually, picking a different window function gives you a different audio file calculation result. In addition to the above window function, there are the following three common window functions:

Hamming window:

$$W_{ham}(n)=\begin{cases}0.54-0.46\cos(\frac{2\pi n}{N}-1) & 0\le n\le N\\ 0 & n\le 0, n\ge N-1\end{cases} \tag{4}$$

Hanning window:

$$\mathrm{W}_{han}(n)=\begin{cases}0.5\left[1-\cos\left(\frac{2\pi n}{N}-1\right)\right] & 0\le n\le N-1\\ 0 & n\le 0, n\ge N-1\end{cases} \tag{5}$$

Blackman Window:

$$W_{bm}(n)=\begin{cases}0.42-0.5\cos(\frac{2\pi n}{N})0.08\cos(\frac{4\pi n}{N}) & 0\le n\le N\\ 0 & n\le 0, n\ge N-1\end{cases} \tag{6}$$

where N are the window lengths of the plus windows.

Based on their expressions, it can be concluded that although rectangular window has smoothness, it has less bandwidth and band leakage occurs. The Hamming window is more suitable for speech signal processing due to its larger bandwidth and removing the effect of Hamming window in terms of smoothness. In addition, the choice of window length is also very important, which directly determines the amplitude variation range of the speech signal, and the choice of the appropriate window length can be a complete reflection of the detailed characteristics of the speech signal, which plays a vital role in the processing of speech signals.

II. C.Speech Feature Extraction

Raw speech waveforms can not be directly recognized and processed, the variable length of the original speech time series signal into a feature vector representation, in order to be directly used as an input to the learning algorithms of intelligent devices, the process is called the feature extraction of speech signals. The process of speech feature extraction is also crucial when using speech signals as input to machine learning algorithms. The feature parameters used for speech recognition should be able to reflect the essential characteristics of the speech signal, and to make the recognition algorithm simpler, the computation process of the feature parameters should be simplified, and the correlation between the parameter components should be reduced as much as possible in the compression process of the data. At present, the commonly used feature parameters are spectrogram, Fbank (filter bank) features and Mel Frequency Cepstrum Coefficient (MFCC). Next, we will introduce three commonly used feature extraction methods for speech signals, which are spectrogram, Fbank (filter bank) features and Mel frequency cepstrum coefficients (MFCC), and this paper adopts the Mel frequency cepstrum coefficients to complete the feature extraction analysis, and then we will conduct a comparative analysis of the three methods, aiming at verifying the priority of the Mel frequency cepstrum coefficients.

II. C. 1) Speech maps

The spectrogram shows through a two-dimensional scale how the intensity of the speech signal varies in different frequency bands due to time variations. First, by processing the symmetric spectrum, we can obtain the frequency spectrum curve with positive frequency axis and stitch the spectrum values of each frame in chronological order. The horizontal coordinate of the spectrogram is time and the vertical coordinate is frequency. In addition, the color depth also affects the spectral value, i.e., the darker the color, the larger the spectral value. Conversely, the lighter the color, the smaller the spectral value, Figure 5 shows the speech spectrogram. Speech spectrogram, also known as time-frequency diagram, is obtained by short-time Fourier transform, which can effectively analyze and study speech information. Compared with a single signal in the time or frequency domain, the spectrogram contains information not only in the time domain but also in the frequency domain. By analyzing the spectrogram, we can obtain the time-domain characteristics of the speech signal and also the frequency-domain characteristics of the speech signal, as well as understand the relationship between the two.

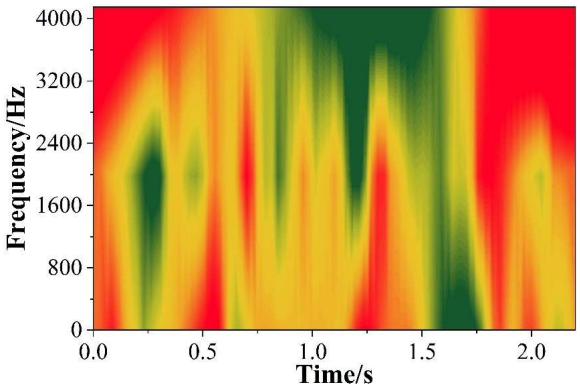


Figure 5: Speech spectrogram

II. C. 2) Fbank Characterization

The Fbank feature extraction process is as follows:

(1) The speech signal is processed through pre-emphasis, frame-splitting and adding Hamming window, while the spectrum of the speech signal can be obtained using Short Time Fourier Transform (STFT).

(2) Calculate the square of the spectrum, i.e., the energy spectrum. By superimposing the energy in each filter band, the k th filter produces the output power spectrum of $X[k]$.

(3) By calculating the output energy of each filter and taking the logarithm of it, the logarithmic power spectrum of the frequency band in question can be obtained. To wit:

$$Y_{Fbank}[k] = \log X[k] \tag{7}$$

Fbank features are essentially logarithmic power spectra, including both low and high frequency information, but compared to spectrogram features, Fbank features are processed through a Mel filter bank in order to compress them better according to the auditory perceptual properties of the human ear, and at the same time suppressing some redundant information that cannot be perceived by the auditory senses.

II. C. 3) MFCC

MFCC is a characteristic parameter that combines the auditory properties of the human ear and the synthetic properties of speech to simulate the human ear to perceive speech signals of different frequencies, and human beings discriminate sound frequencies in a linear relationship that increases or decreases exponentially, as if taking logarithmic operations [20]. The frequency range of sound signals is between 20-20,000 Hz to be perceived by the human ear, and the ease with which different audio signal frequencies within this range are perceived by humans varies. Frequency in the 1000Hz below the sound, the human perception ability and the frequency of the sound into a linear relationship. For audio signals with a frequency greater than 1000Hz, the human perceptual ability is close to a logarithmic relationship with the size of the sound frequency. According to this nature of the human ear to the sound frequency, the Mel cepstrum coefficient is proposed, Mel frequency and the actual frequency of the relationship shown in equation (8):

$$Mel(f) = 2595 \cdot \lg\left(1 + \frac{f}{700}\right) \tag{8}$$

where f is the actual linear frequency and $Mel(f)$ is the Mel cepstrum frequency, both in Hz.

The specific extraction steps of MFCC are as follows:

In the first step, preprocessing operation is performed on the input audio data.

In the second step, Fast Fourier Transform (FFT), FFT operation is shown in equation (9):

$$X(i,k) = FFT[x_i(m)] \tag{9}$$

In Equation (9), i is the speech signal, k is the frequency, $X(i,k)$ is the frequency domain signal, and $x_i(m)$ is the time domain signal. Subsequently, the spectral line energy of the speech signal is calculated:

$$E(i,k) = \left[X(i,k)\right]^2 \tag{10}$$

In the third step, the energy of the Mel filter bank is calculated:

$$S(i,m) = \sum_{k=0}^{N-1} E(i,k) H_m(k), 0 \le m \le M \tag{11}$$

In Equation (10), $E(ik)$ is the spectral line energy. In Equation (11), $H_m(k)$ is the frequency domain response of the Mel filter bank. Multiplying and weighting $E(ik)$ and $H_m(k)$ gives $S(i,m)$, which is the total energy after passing through the Mel filter bank.

Mel filters are similar to triangular filters and are denoted by $H_m(k)$, the Mel filter bank consists of multiple $H_m(k)$ combinations, where $0 \le m \le M$, M is the number of filters. $f(m)$ is the center frequency, and Eq. (12) is the transfer function of the bandpass filter. To wit:

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \le k \le f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{12}$$

In the fourth step, the discrete cosine transform (DCT), i.e., the inverse spectral operation. The calculation process is shown in equation (13):

$$MFCC(i,n)=\sqrt{\frac{2}{M}}\sum_{m=0}^{M-1}\log[S(i,m)]\cos\left(\frac{\pi n(2m-1)}{2M}\right),n=0,\cdots,2M-1 \tag{13}$$

II. D.Algorithm Improvement Design

After completing the audio pre-processing operation and speech feature extraction operation above, the next step is to utilize the DCNN-CTC model and ASKCC network to complete the design of speech recognition algorithm improvement, in order to enhance the results of students' English listening training in the context of intelligent education.

II. D. 1) DCNN-CTC modeling

The model accuracy is improved by deepening the layers of the model network to learn more advanced feature information. Figure 6 shows the structure diagram of the acoustic model implemented using deep convolutional neural network, which is mainly composed of two parts: deep convolutional neural network (DCNN) and connection timing classification (CTC). The details are as follows.

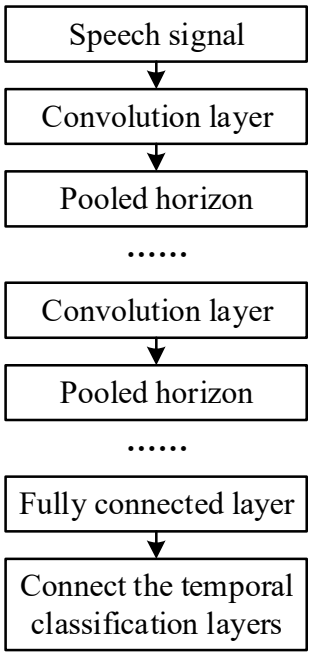


Figure 6: The structure of DCNN-CTC

The convolutional layer is responsible for capturing local feature information at different locations, while the pooling layer applies maximum pooling for feature extraction [21]. To expand the field of view and get the global feature information by downsampling, the network has fewer parameters and is simple to train, which is formulated as follows:

$$y^{(i)}=\sigma(W^{(i)}*y^{(i-1)}+b^{(i)}) \tag{14}$$

$$y^{(i+1)}=f_{\max pool}(y^{(i)}) \tag{15}$$

where σ denotes the nonlinear operation in the convolutional layer, $W^{(i)}$ denotes the weight parameter in the layer, $b^{(i)}$ denotes the bias in the layer, $y^{(i)}$ denotes the output of the i th layer, and $f_{\max pool}$ denotes the max pooling operation.

It is assumed that each output variable occurs conditionally independent when the input variables are determined. When a speech feature A of length T is input to the acoustic model, the conditional probability p that the output is a correctly decoded path π can be computed from equation (16):

$$p(\pi \mid A) = \prod_{t=1}^{T} p(\pi_t \mid A) \tag{16}$$

CTC introduces a "blank" label to model the gaps, overlaps, etc. in speech. As shown in Eq. (17), there are multiple decoding paths under one label. Namely:

$$\left.\begin{array}{c} Y(z,-,-,c,-c,l) \\ Y(zz,-,-,cc,-,l) \\ Y(z,c,-,-,-,l) \\ \ldots \end{array}\right\} \tag{17}$$

where "-" denotes the "blank" label, and Y denotes a mapping function that maps to the same target label (z,c,l) no matter where "blank" is inserted. The CTC model will summarize the probabilities of all potential paths with the following formula $p(L \mid A)$:

$$p(L \mid A) = \sum p(\pi_i \mid A) \tag{18}$$

where π_i denotes each possible path and L denotes the target label, and then the CTC loss value is calculated by Equation (19):

$$Loss_{cic} = -\ln p(L \mid A) \tag{19}$$

The model is fine-tuned in the training phase according to the loss value drop to achieve a better fit. The model integrates the labels corresponding to the maximum value of $p(L \mid A)$ in the best path decoding stage to get the final output sequence.

II. D. 2) ASKCC network

Aiming at the three problems described above, this paper proposes an ASKCC network:

(1) Taking advantage of the SKNet network, on the time axis, it uses a multi-size convolution kernel to compute the feature map, fuses the convolution results, and realizes the normalization of phoneme features. It can realize the reasonable allocation of weights to the feature information and avoid the redundancy of feature information.

(2) In this paper, the causality of the convolution layer is realized by modifying the filling method of the convolution layer to VALID and artificially filling two frames of zeros in front of the input feature data, which well solves the information leakage problem in the above.

(3) In this paper, according to the model structure, an attention mechanism (ASKCC) is well introduced, which effectively extracts more key features and improves the model accuracy, and constitutes a new acoustic model (ASKCC-DCNN-CTC), and the structure of ASKCC-DCNN-CTC is shown in Fig. 7.

II. D. 3) Causal Convolution

Convolutional neural networks are widely used in image-video processing, speech enhancement, and other fields due to their weight sharing, local connectivity, and powerful modeling capabilities. The dataset qualities significantly affect the prediction accuracy of neural network models. The traditional CNN will cause problems such as information leakage when performing convolutional computation due to the utilization of future frames data, which in turn affects the prediction accuracy of the model in the prediction stage. When it is also necessary to input 5 frames of feature data, and the convolution size is (3×3), this paper realizes the causality of the convolution layer by modifying the padding of the convolution layer to VALID and artificially padding two frames of zeros in front of the input feature data, which shows that at the moment of t_1 only t_0, t_1 and t_{-1} moments, and does not utilize data from future frames, which well solves the information leakage problem mentioned above.

II. D. 4) SKNet-based multiscale convolution

X and V denote the input data and the feature data after adaptive fusion. w, h, and c represent the three dimensional information of the feature map, which denote the width of the feature map, the height of the feature map, and the number of channels of the feature map, respectively. $\tilde{F}$, $F_{gp}(U)$ and F_{fc} denote some nonlinear operations of the SKNet network species, which are computed as follows:

$$\tilde{F} = conv(X, f, same, k = (3,1)) \tag{20}$$

$$\tilde{F} = conv(X, f, same, k = (5,1)) \tag{21}$$

$$F_{gp}(U)=\frac{1}{W\times H}\sum_{i=1}^{H}\sum_{j=1}^{W}U(i,j) \tag{22}$$

$$F_{fc}=\delta(B(W_s)) \tag{23}$$

$$softmax=\frac{e^{x_i}}{\sum_{j=1}^{N}e^{x_j}} \tag{24}$$

where $conv(X,f,same,k=(s,s))$ denotes the input of X to a convolutional layer with channel number f, convolutional kernel size (s,s), and padding mode same for computation, δ denotes the RELU activation function, B denotes the Batch Noramlization normalization operation, W_s denotes a new feature obtained by shortening once more the features generated after the $F_{gp}(U)$ operation, and softmax denotes a nonlinear operation whose input is a sequence X of length N.

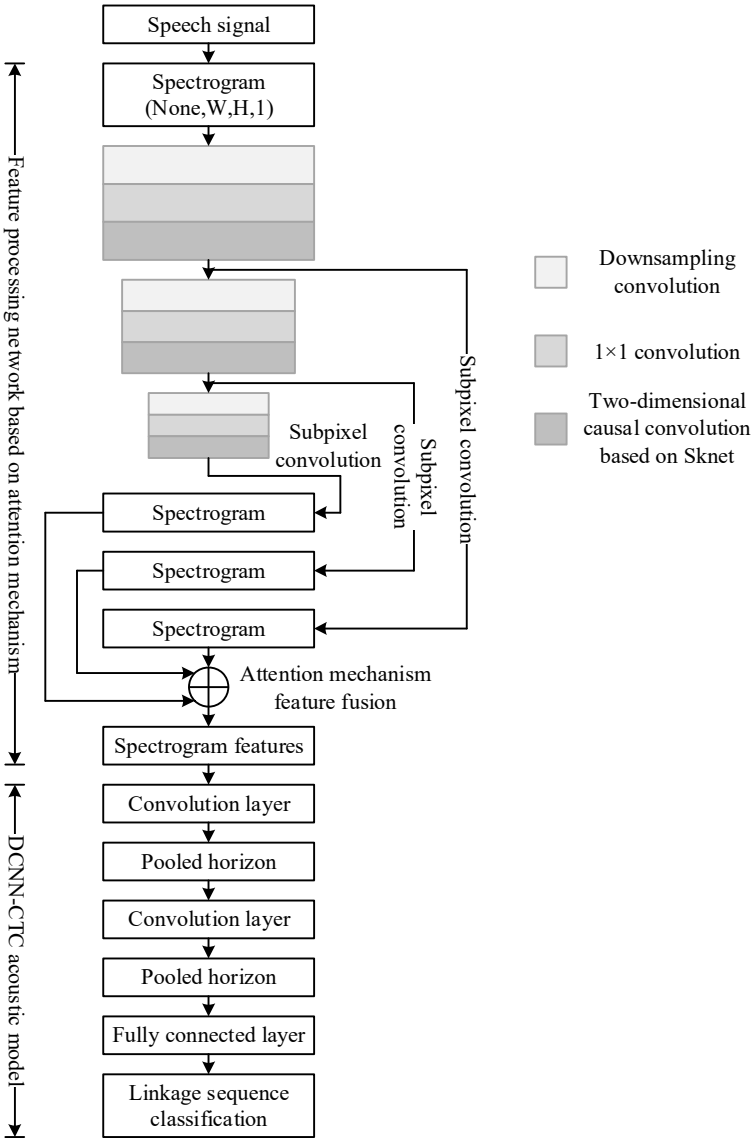


Figure 7: ASKCC-DCNN-CTC Structure Diagram

II. D. 5) Attention-based mechanism feature fusion

In order to pay more attention to the features that play a decisive role in the prediction results of the acoustic model and reduce the problem of feature information redundancy, an attention mechanism is introduced to reasonably assign weights to the three amplified feature maps and realize the fusion. The calculation formula is as follows:

$$D = A + B + X \tag{25}$$

$$E(D) = \frac{1}{W \times H} \sum_{i=1}^{H} \sum_{j=1}^{W} D(i, j) \tag{26}$$

$$F_{fc} = \delta(B_n(W_s)) \tag{27}$$

$$F_{mul}(V) = \rho(V) \tag{28}$$

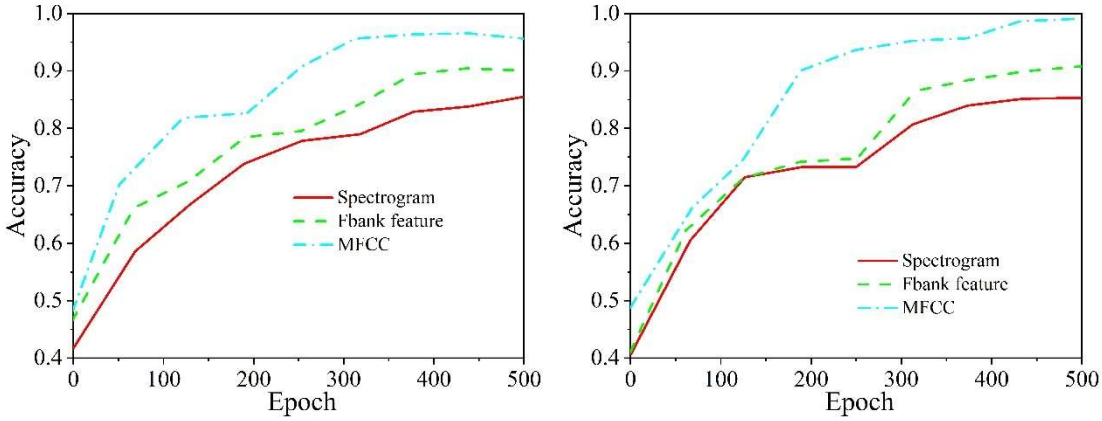
δ denotes the RELU activation function, B_n denotes the normalization operation, and ρ denotes an operation that multiplies the elements of a tensor. Firstly, the three features A, B and X output from sub-pixel convolution are accumulated and summed to get D, secondly, the weights of each channel are calculated E, and then the feature dimensions are enlarged and compressed again to make the network fit the correlation between the channels better, and finally, the outputs of the three non-linear operations are multiplied by A, B and X respectively to get the feature map O that occupies more key feature information of the feature map O.

III. Speech recognition algorithm improvement empirical research analysis

III. A. Analysis of Feature Extraction in the Process of English Listening Training

III. A. 1) MFCC algorithm feature extraction performance test

In order to verify the superiority of the MFCC algorithm in the field of speech feature extraction for intelligent translation robots, the experiments introduced the speech spectrogram, Fbank features to compare the experiments with the MFCC algorithm. The algorithm is trained with ATIMIT dataset and Librispeech dataset, and the experimental results are shown in Fig. 8, where (a)~(b) are ATIMIT dataset and Librispeech dataset, respectively. From Fig. 8(a), it can be seen that in the ATIMIT dataset, the proposed MFCC algorithm performs the best, and the accuracy change curve tends to converge after 400 iterations, and finally converges to 0.935, and the performance of the spectrogram and the Fbank features are not much different, and the accuracy converges to about 0.7~0.8. From Figure 8(b), it can be seen that in the Librispeech dataset, the accuracy of the proposed MFCC algorithm does not change much. In contrast, the accuracy of Speech Spectrogram and Fbank features varies more, which implies that the research-proposed MFCC feature extraction scheme has good generalization ability.



(a)Data set ATIMIT (b)Data set Librispeech

Figure 8: Feature extraction performance test

The study utilizes DCT to extract MFCC features, and in order to verify the superiority of this scheme, the experiment introduces Discrete Fourier Transform (DFT), Discrete Sine Transform (DST) for comparison, and the experimental results are shown in Fig. 9. Based on the data presented in Fig. 9, it is easy to find that the accuracy

of the newly proposed speech feature extraction algorithm presents significant differences when dealing with different languages. Specifically, the algorithm performs relatively poorly on British English feature extraction, with a low accuracy rate. On the contrary, when dealing with American English, its extraction accuracy reaches a very satisfactory level. Especially for the DCT method, its accuracy is as high as about 96.82%, showing excellent performance. Meanwhile, the accuracy of DST and DFT methods also reached 95.73% and 94.87%, further verifying the effectiveness of the algorithm in English speech feature extraction.

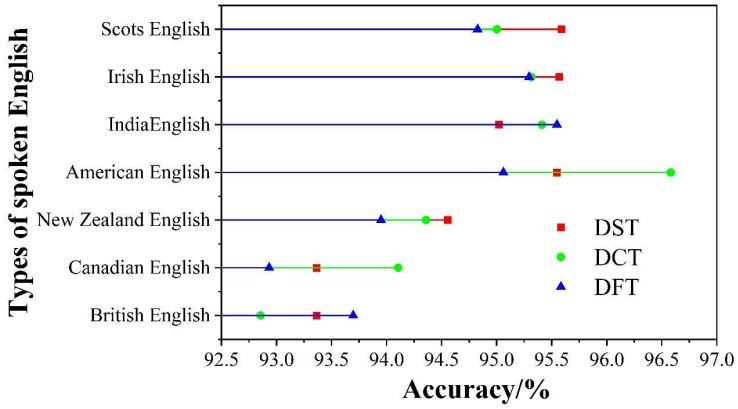
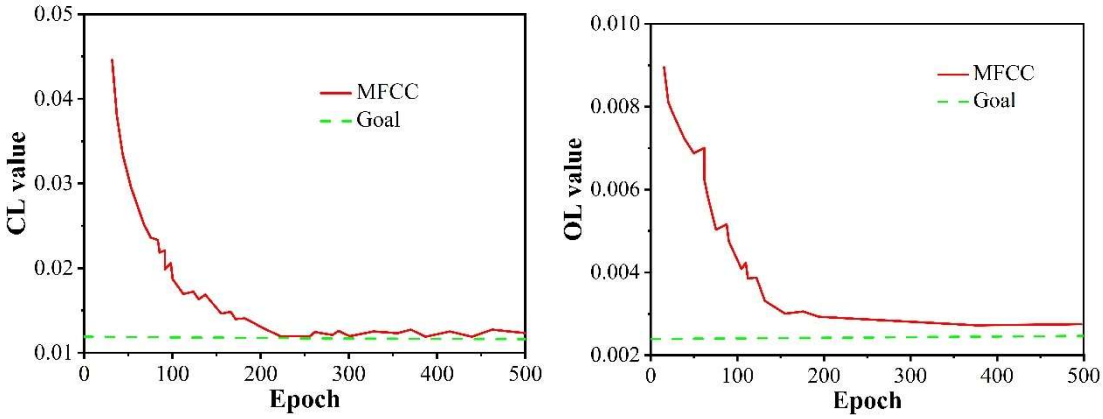


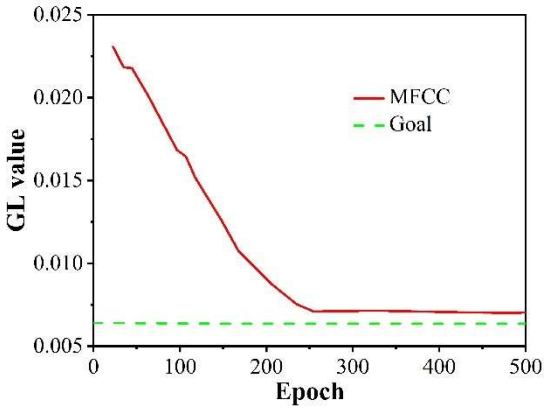
Figure 9: The accuracy rate of feature extraction in oral English

III. A. 2) MFCC algorithm application-based testing

In order to understand the various aspects of the performance of the smart English translator equipped with the MFCC model, the study recruited 100 volunteers from the College of English and conducted data collection, which yielded 10,000 pieces of data, and the model was trained with this dataset. The classification loss curve (CL), the generalized intersection and merger ratio loss curve (GL), and the target loss curve (OL) were used as evaluation indexes, and the experimental results are shown in Figure 10. From Fig. 10(a), it can be clearly seen that at the early stage of training, the classification loss declines rapidly, and with the continuous advancement of training rounds, the curve continues to decline and stabilize, and finally converges successfully to 0.0128 after 215 rounds of training, which is a result that fully reflects the accuracy and high efficiency of the algorithm in the classification task. From Fig. 10(b), we can see that the loss curve of the generalized intersection and merger ratio of the MFCC algorithm also shows a fast decreasing trend at the early stage of training, and after 158 rounds of detailed training, it converges to 0.00321. From Fig. 10(c), it can be seen that the MFCC algorithm has a much better performance in terms of the target loss curve, achieving an excellent convergence value of 0.00783 in only 256 rounds of training, demonstrating the algorithm's ability in target localization.



(a)CL of MFCC (b)OL of MFCC



(c)GL of MFCC

Figure 10: CL, GL, OL change curve of MFCC algorithm

In order to further verify the superiority of the proposed algorithm, the experiment introduces the WOA-PSO algorithm as a comparative experiment with MSE as the evaluation criterion, and Figure 11 shows the MSE scatter plots of each data point of the two schemes.The mean square error (MSE) of the MFCC algorithm mainly concentrates in the range from 0.015 to 0.025, which suggests that the algorithm has a higher degree of accuracy and stability in the English translation task. In contrast, the performance of the WOA-PSO algorithm is relatively poor, and its MSE is mainly distributed between 0.025 and 0.055, this result reflects that the algorithm may have large errors and uncertainties in the process of feature extraction.

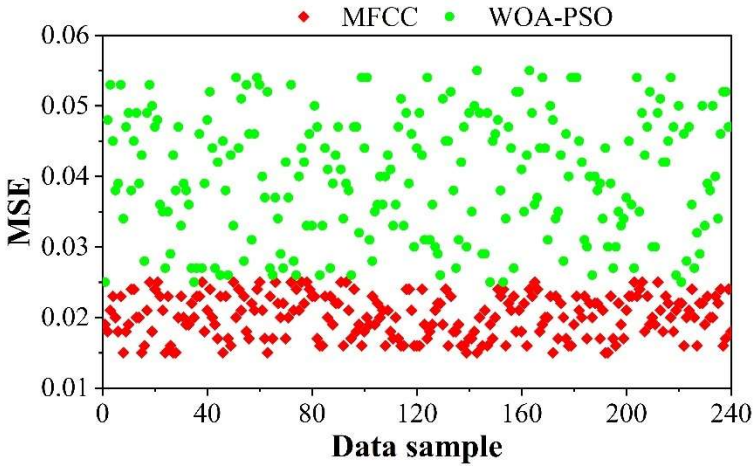


Figure 11: MSE scatter plots for each data point of the two schemes

III. B. Validation analysis of ASKCC-DCNN-CTC based recognition algorithm

III. B. 1) Experimental setup

In this paper, MFCC is used as the input of the network, the frame length is set to 8ms, the frame shift is set to 7ms, and the mel spectral dimension is set to 256 when extracting the speech features, and the dataset contains a total of 5 segments of audio, and this paper carries out the 5-fold cross-validation experiments with the unit of segments. In addition, because of the unequal length of the audio in the dataset, the long paper in the data processing part of the data length uniformly 6.8s of audio for the complementary zero operation, longer than 6.8s of audio truncated to 6.8s.

III. B. 2) Evaluation indicators

In this paper, the confusion matrix is used to evaluate the classification performance, and the confusion matrix includes 2 specific evaluation metrics, unweighted recall (UA) and weighted accuracy (WA). The specific calculations are as follows:

$$UA = \frac{\sum_{i=1}^{4} \frac{aa_{ii}}{\sum_{j=1}^{4} a_{ij}}}{4} \tag{29}$$

$$WA = \frac{\sum_{i=1}^{n} aa_{ii}}{\sum_{i=1}^{4}\sum_{j=1}^{4} a_{ij}} \tag{30}$$

III. B. 3) Experimental results

Firstly, in this paper, when computing the attention mechanism, the global pooling tries to adopt average pooling and maximum pooling respectively, and compares the evaluation indexes with the DCNN-CTC model when the attention mechanism is not added, and the results of the attention mechanism cross-validation are shown in Table 1. After the introduction of the attention mechanism (ASKCC), regardless of whether global average pooling or global maximum pooling is used, there is a significant performance improvement compared to the DCNN-CTC model when the attention mechanism is not introduced, in which the effect of channel attention using global average pooling is the most obvious, with the UA achieving an improvement of 2.98% and the WA achieving an improvement of 4.4%.

Table 1: Results of cross-validation of attention mechanisms

Model	UA/%	WA/%
Unintroduced attention mechanism	65.87	64.53
Global average pooling	68.85	65.93
Global maximum pooling	66.54	65.68

In order to enhance the performance ability of the network and retain the original feature distribution, this paper introduces the SKNet structure on the basis of the attention mechanism and adjusts the network weighting, and the cross-validation results of the attention mechanism with the introduction of SKNet are shown in Table 2. The ASKCC-DCNN-CTC model with the introduction of the SKNet structure has the highest speech recognition accuracy than the ASKCC-DCNN-CTC model without the introduction of the SKNet structure, in which the ASKCC-DCNN-CTC model with the global average pooling and the weighting of 2 achieves an improvement of 2.96% for the UA, and 4.44% for the WA, compared to the ASKCC-DCNN-CTC model without the introduction of the SKNet structure. 4.44% improvement.

Table 2: Introduction of SKNet's attention mechanism cross-validation results

Model	Network gravity	UA/%	WA/%
Global average pooling	1	68.85	68.94
Global maximum pooling	1	68.64	65.87
Global average pooling	2	68.83	68.97
Global maximum pooling	2	68.85	65.63

Overall, this paper introduces the attention mechanism into the DCNN-CTC model to realize the attention to different channels in the process of network learning. 2 ways of global average pooling and global maximum pooling are tried in calculating the attention, and the residual structure is introduced to explore the different kinds of attention methods. The final experimental results show that the accuracy of the speech recognition algorithm after the introduction of the attention mechanism is significantly improved, which indicates that the different channels of the DCNN output contain different speech features, and the performance of the recognition algorithm based on the DCNN-CTC model can be improved after the introduction of the attention mechanism, and at the same time, it plays a very good role in the improvement of the above-described three problems, and it can well satisfy the needs of students' English listening in the context of wisdom training needs.

IV. Conclusion

This paper first describes the problems of current speech recognition algorithms, for the problem, proposes a recognition algorithm based on ASKCC-DCNN-CTC, and verifies the empirical analysis of the algorithm.

(1) After 400 iterations of training, the MFCC feature extraction algorithm achieves a feature extraction accuracy of 0.935, which is more outstanding than the spectrogram algorithm and the Fbank feature algorithm, and verifies the priority of the MFCC feature extraction algorithm.

(2) On the basis of the speech recognition algorithm based on the DCNN-CTC model, after the introduction of the attention mechanism (ASKCC), it is found that the unweighted recall (UA) and the weighted precision are improved, with the values of 2.98% and 4.4%, respectively, which indicates that the attention mechanism has an improvement effect on the speech recognition algorithm based on the DCNN-CTC model, so as to make it better serve the students of colleges and universities. English listening training.

References

[1] Guo, H. (2022). Research on the Current Situation and Countermeasures of English Listening Teaching Based on Multimedia Intelligent-Embedded Processor. Journal of Sensors, 2022(1), 9722209.

[2] Zhang, J. (2023). A Practical Study on the Blended Teaching Model of English Listening and Speaking Based on the Smart Language Laboratory. The Educational Review, USA, 7(9), 1289-1294.

[3] Wei, L., & Liu, B. (2022). Smart Classroom College English Listening Teaching System Based on Virtual Environment Technology. Journal of Cases on Information Technology (JCIT), 24(5), 1-18.

[4] El-Sourani, A. I., Keshta, A. S., & Aqel, M. S. (2021). The Effectiveness of Educational Environment Based on Smart Learning in Developing English Language Listening Skill among IUG Female Learners. IUG Journal of Educational & Psychological Studies, 29(6).

[5] Amakhina, S., Dmitrieva, N., & Timokhina, E. (2023, November). Improving speaking and listening skills: An Educational eco-system for foreign languages teaching in higher education. In International Conference on Professional Culture of the Specialist of the Future (pp. 402-413). Cham: Springer Nature Switzerland.

[6] Gebregziabher, H. (2024). IMPROVING ENGLISH LISTENING AND SPEAKING ABILITY BASED ON ARTIFICIAL INTELLIGENCE WIRELESS NETWORK. Machine Intelligence Research, 18(1), 155-166.

[7] Dennis, N. K. (2024). Using AI-Powered Speech Recognition Technology to Improve English Pronunciation and Speaking Skills. IAFOR Journal of Education, 12(2), 107-126.

[8] Hu, N. (2024). English listening and speaking ability improvement strategy from Artificial Intelligence wireless network. Wireless Networks, 1-10.

[9] Yuniarti, F., Wulandari, F., & Rakhmawati, D. (2024). Personalized Listening Practice: Integreting AI (Chat GPT) and (Voxbox) Native Speaker Input in Language Learning. INOVISH JOURNAL, 9(2).

[10] Sahito, J. K. M., Panwar, A. H., & Ramzan, I. (2025). EXPLORING THE IMPACT OF ARTIFICIAL INTELLIGENCE (AI) ON THE LISTENING SKILLS OF ENGLISH AS A SECOND LANGUAGE (ESL) LEARNERS. Journal of Applied Linguistics and TESOL (JALT), 8(1), 1059-1067.

[11] Xiao, Y. (2025). The impact of AI-driven speech recognition on EFL listening comprehension, flow experience, and anxiety: a randomized controlled trial. Humanities and Social Sciences Communications, 12(1), 1-14.

[12] Ran, D., Yingli, W., & Haoxin, Q. (2021). Artificial intelligence speech recognition model for correcting spoken English teaching. Journal of Intelligent & Fuzzy Systems, 40(2), 3513-3524.

[13] Li, D., Yang, M., Lyu, T., Li, B., Zhao, Y., & Qin, S. (2024, November). Application of Automatic Speech Recognition Theory in Improving Pronunciation and Listening Skills for EFL Learners. In 2024 International Conference on Cyber-Physical Social Intelligence (ICCSI) (pp. 1-6). IEEE.

[14] Wang, Y. (2023). An English listening and speaking ability training system based on binary decision tree. International Journal of Continuing Engineering Education and Life Long Learning, 33(2-3), 313-325.

[15] Mirzaei, M. S., Meshgi, K., & Kawahara, T. (2018). Exploiting automatic speech recognition errors to enhance partial and synchronized caption for facilitating second language listening. Computer Speech & Language, 49, 17-36.

[16] Liu, J. (2024, May). Application of Machine Learning in Adaptive English Listening Training System. In The World Conference on Intelligent and 3D Technologies (pp. 485-496). Singapore: Springer Nature Singapore.

[17] Liu, Y., & Quan, Q. (2022). AI recognition method of pronunciation errors in oral English speech with the help of big data for personalized learning. Journal of Information & Knowledge Management, 21(Supp02), 2240028.

[18] Jingning, L. (2024). Speech recognition based on mobile sensor networks application in English education intelligent assisted learning system. Measurement: Sensors, 32, 101084.

[19] Wenbo Zhang,Xuefeng Xie,Yanling Du & Dongmei Huang. (2024). Speech preprocessing and enhancement based on joint time domain and time-frequency domain analysis. The Journal of the Acoustical Society of America,155(6),3580-3588.

[20] Mahbubeh Bahreini,Ramin Barati & Abbas Kamali. (2025). Cardiac sound classification using a hybrid approach: MFCC-based feature fusion and CNN deep features. EURASIP Journal on Advances in Signal Processing,2025(1),2-2.

[21] Sen Lin Xie,Anfeng Hu,Meihui Wang,Zhi Rong Xiao,Tang Li & Chi Wang. (2025). 1DCNN-based prediction methods for subsequent settlement of subgrade with limited monitoring data. European Journal of Environmental and Civil Engineering,29(4),759-784.