Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

A Study of Teaching Quality Improvement in English Listening Teaching in the Context of Speech Recognition Technology

Di Wang[1], Lili Zhang[1,✉]

[1] *Section of Fundamentals, Shijiazhuang Institute of Railway Technology, Shijiazhuang, Hebei, 050000, China*

ABSTRACT
At present, most English learners spend much less time listening to English than reading it. Most of the language knowledge is acquired through visual channels rather than auditory channels, thus the language knowledge does not form corresponding auditory images in the mind, so the phenomenon of reading but not understanding occurs. Aiming at this kind of problem, this paper tries to explore the role of speech recognition in improving the quality of English teaching by combining it with this technology. The article first recognizes English spoken speech features based on Mel's frequency cepstrum feature parameters and deep belief network, then expands the number of speech features from both time and frequency by means of distortion and masking, and designs the encoder part by combining 2D convolutional neural networks and GRUs, and finally models the local and temporal information in the speech features to realize the recognition of English speech. And thus establish a new model of English listening teaching. As verified by the dataset, the method in this paper can accurately recognize speech features of different emotions, and the recognition effect is better than other models of the same type. In addition, an equivalence study between the proposed teaching model and the traditional teaching model was conducted with 70 foreign students in a university. It was found that the mean value of the total scores of the candidates in the group of the teaching model proposed in this paper was 0.26 points higher than the mean value of the total scores of the candidates in the traditional group, among which, the mean value of the listening scores of the candidates in the group of the teaching model proposed in this paper was 0.1 points higher than the mean value of the listening scores of the candidates in the traditional group.

Keywords: speech recognition, speech features, convolutional neural network, GRU, listening teaching

1. Introduction

With the deepening of economic globalization, the communication between China and the rest of the world is getting closer and closer, and English, as the most widely used speech in the world, has become one of the important tools for people's daily life and work. As college students, they should have the ability to use this tool flexibly [1-3]. At present, all universities offer English courses in

✉ Corresponding author.
E-mail address: zhang_sirt_2003@163.com (L. Zhang).
Received 10 January 2024; Revised 20 February 2024; Accepted 10 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-238

freshman and sophomore years as a key initiative to improve college students' English proficiency. Many college students' English ability has been practiced, and the level of English teaching has been effectively improved, but the overall level needs to be further improved, especially the ability to speak and listen to English is relatively weak. Many college students are still difficult to open their mouths to speak and understand English [4-6].

Speech recognition is the process of automatically transforming speech signals into corresponding text or commands through computers and software. The essence of speech recognition is a kind of pattern recognition based on the parameters of speech features, i.e., through learning, the system is able to categorize the input speech according to a certain pattern, and then find out the best matching result based on the decision criterion. With the development of computer technology, many computer-assisted English learning systems have been used in large numbers [7-9]. And based on the use of speech recognition technology, the need to provide individualized practice and real-time feedback to each student becomes a reality [10].

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 the aspect of listening, speech recognition has made great progress, and the accuracy rate is constantly improving [11-12]. On the speaking side, it is even simpler. Speech synthesis technology has become so mature that, except for minor problems such as sometimes having intonation or breaking sentences, the effect of automatic speech synthesis based on text can be compared with that of a real person. Translators using deep learning, on the other hand, have demonstrated the ability of the reading and writing side, before the emergence of the attention mechanism, RNN used to achieve a number of impressive results, and after the emergence of the attention mechanism and BERT, the level of translation has been further improved [13-14]. Not only that, AI can also be used to make texts and even write poems, and deep learning neural networks that can look at pictures and make texts have also appeared. All these results show that the language skills that used to require boring learning and long-term accumulation can be easily obtained using artificial intelligence. This will bring about a sea change in English language teaching [15-16].

However, despite the fruitful achievements of AI, there are still differences in its performance in different fields, especially in machine translation, which is still difficult to replace human translation in all fields because language can be used to express knowledge in different fields of specialization, and the use of language may also be influenced by many factors such as allusion, allusion, tendency, and context [17-19]. This problem should not change in the foreseeable future (decades) either. In contrast, speech recognition performs slightly better, with errors coming mainly from background noise. On occasions where the signal-to-noise ratio is high, the correctness of speech recognition can be quite high. Considering that Chinese students' English proficiency has a prominent feature of being good at reading and writing but poor at listening, a new possibility arises in situations where communication with British and American people is required, where speech recognition software converts English speech into English text and displays it to the Chinese, thus facilitating communication in one direction [20-22]. Of course, speech recognition does not help in reverse communication (it is impossible for foreigners to read Chinese text), but another feature of linguistic communication is that native speakers will easily understand what non-native speakers say. So the reverse communication barrier will be much smaller. Based on these characteristics, a completely new mode of foreign communication will emerge [23-24]. With the continuous development of social internationalization, in college English teaching, colleges and universities pay more and more attention to the cultivation of students' listening and speaking ability, but the actual teaching effect is not optimistic, therefore, it is necessary to explore the application of speech recognition technology in college English listening teaching to improve students' English listening and speaking ability [25].

English listening teaching is an important part of English learning, and English listening significantly affects students' English communication, so how to improve English listening has always been a key

concern in the field of English teaching. Some of these scholars have revealed the status quo of English listening learning and the obstacles encountered through theoretical analysis and teaching research, and put forward some suggestions, mainly descriptive research, in a targeted manner. Literature [26] describes the status quo of English listening learning, argues that English listening learning is a receptive skill that focuses on the active process, and records the difficulties and obstacles encountered by students in the process of English listening learning and puts forward suggestions in a targeted manner. Literature [27] discusses the process of English learning, due to the limited real application of English, so the students' English listening skills improvement is very limited, so it focuses on analyzing the root causes of the students' English listening problems, and at the same time, it summarizes the opinions of experts in English listening teaching and the principles of English listening teaching in the present time, which makes a positive contribution to the further reform and optimization of English listening teaching. And the most scholars pay attention to improving English listening teaching methods and approaches, which include multimedia technology teaching platform, English songs, and scaffolding strategies for authentic listening materials. Based on the theory of multimedia English teaching, literature [28] designed an English listening teaching platform in colleges and universities with multimedia technology as the core logic, and clarified the demand for listening teaching through the questionnaire survey method, optimized the English listening teaching platform, and finally confirmed the feasibility of the constructed English listening teaching platform based on the teaching experiments. Literature [29] qualitatively researched the role of English songs in improving students' English listening skills based on the questionnaire survey method, and confirmed the positive role of English songs in promoting the cultivation of students' English listening skills based on the research results. Literature [30] emphasized that listening comprehension ability has a rising status in language teaching, but it is very difficult for students to improve their English listening skills, so it argues and analyzes the scaffolding strategy of applying authentic listening materials in listening teaching, aiming at helping English learners to improve their English listening skills. Meanwhile, some other scholars conducted research on the evaluation of English listening teaching methods, which generally used the questionnaire method. Literature [31] combined the descriptive qualitative analysis method to explore students' online English listening learning mode based on the information data from the questionnaire survey and the final assignments of the English course, and the study pointed out that students' online and offline media listening learning modes accounted for half of the students' online and offline media listening learning modes, in which 2/3 of the listening learning methods chose audio-visual methods, and 1/6 each of the scientific method and community-based language learning. Literature [32] elaborated that listening is an an important skill for second language learning, so a systematic review of relevant studies on listening skill development in second language classroom teaching provides certain optimization suggestions and guidance for second language listening teaching.

Speech recognition technology plays an important role in promoting the enhancement of spoken English teaching. Speech recognition technology is mainly applied to the recognition of spoken English and the correction of spoken English, which has an auxiliary role in the cultivation and enhancement of English learning students' listening skills. Literature [33] proposes an English speaking correction model based on speech recognition technology, and through simulation experiments, it is verified that the model can effectively filter speech noise, accurately recognize speech emotion and English accent, and has a high response time. Literature [34] describes the wide application of speech recognition technology in English teaching, and based on the key segmentation technology in speech recognition technology, it conceptualizes a mindful Chinese segmentation strategy to adapt the advantages of probabilistic statistics, string matching, and semantic understanding to achieve the optimal segmentation result, which has higher speech recognition accuracy and efficiency compared with the traditional method. Literature [35] introduces new methods of generative adversarial networks (GANs) and optimized edge computing (OEC) framework to optimize speech recognition technology, and uses

it as the basis to optimize English listening teaching in colleges and universities, which, combined with a large number of experiments and case studies, confirms that the above optimization scheme has a better pronunciation accuracy and a good real-time interactive feedback experience compared with the traditional speech recognition technology listening teaching scheme. The optimization program of the above mentioned program has better pronunciation accuracy and good real-time interactive feedback experience than the traditional speech recognition technology listening teaching program. Literature [36] conducted language teaching experiments to evaluate the performance of three automatic speech recognition (ASR) software, and confirmed that these three software can effectively assess the dictation level of English learners, which side by side proves that these speech recognition software are feasible and reliable. Literature [37] envisioned a neural network speech recognition algorithm with deep convolutional neural network as the underlying architecture to address the accuracy and efficiency of speech recognition technology in language learning, and based on the performance tests and experiments, it was confirmed that the proposed speech recognition algorithm can support language learners in correcting their pronunciation and improving their English speaking skills.

The article firstly gives an introduction to data preprocessing and feature extraction in speech recognition tasks, applies deep learning techniques to English speech recognition, and adopts Meier frequency cepstrum feature parameters based on the human ear hearing model and deep belief network to establish a speech recognition model. Then the speech recognition feature enhancement part is described in detail, and the feature enhancement module is proposed, which expands the speech features quantitatively from both time and frequency by means of distortion and masking, so that the model can learn better parameters with more data. Then a combination of 2D convolutional neural network and GRU coding is designed in the encoder part to model the local and temporal information of speech features. Subsequently, the effectiveness of the speech recognition model proposed in this paper is shown through experiments. Finally, an equivalence study of traditional teaching and this teaching model was conducted with 70 foreign students in a university to test the effectiveness of the method proposed in this paper.

2. Research on Speech Recognition Technology in English Listening Teaching

2.1. Speech signal preprocessing and feature extraction

2.1.1. Speech recognition process. The general process of speech recognition is shown in Figure 1. First, the voice analog signal is digitized using the PC's sound card and the voice signal is captured. According to the Nyquist sampling theorem: in the process of performing analog/digital signal conversion, when the sampling frequency $f_{s\_\max}$ is greater than two times the highest frequency $F_{\max}$ in the signal, as shown in equation (1):

$$f_{s\_\max} \geq 2 * F_{\max} \tag{1}$$

Then the digital signal after sampling can express the effective information in the original speech signal more completely. Given that the frequency of normal human speech is generally in the range of 40~4000 Hz, this paper sets the sampling frequency to 8 KHz. Then, the resulting speech signal is preprocessed, and the preprocessing includes units such as pre-emphasis, frame-splitting, windowing and endpoint detection. Then, the feature parameters of the pre-processed speech signal are extracted [38]. Finally, the speech feature parameters can be selected for model training or pattern matching.

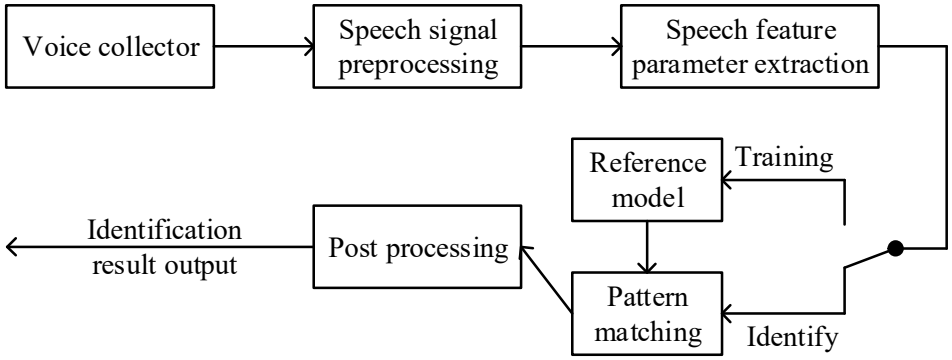

Fig. 1. Speech recognition process

2.1.2. Speech signal preprocessing:

1) Pre-emphasis. The high-frequency end of the average power spectrum of a speech signal decays by 6 dB/oct (octave) above about 800 Hz due to the influence of the mouth and nose radiation and the sound gate excitation. Therefore, before analyzing the speech signal, a 6dB/oct high-frequency boosting pre-emphasis digital filter is generally used to boost the high-frequency portion of the speech signal, so that the spectrum of the speech signal becomes flat, and the same signal-to-noise ratio can be used to find the spectrum of the entire frequency band from low to high frequencies. The filter response function is shown in equation (2):

$$H(z)=1-\alpha z^{-1}, 0.9\le\alpha\le 1.0 \tag{2}$$

where α is the pre-emphasis coefficient, which is taken as 0.9375 in this paper.In this way, the result $y(n)$ after pre-emphasis processing can be represented by the input speech signal $x(n)$ as:

$$y(n) = x(n) - \alpha x(n-1) \tag{3}$$

2) Framing. The analysis and processing of voice signals is usually based on “short-time”, that is, the use of “short-time analysis”, the voice signal stream is divided into frames. The general number of frames per second are

$$Frames\ per\ \sec ond = \frac{1}{t}(0.01 < t < 0.03) \tag{4}$$

3) Windowing. In order to strengthen the speech waveform near sample n and weaken the rest of the waveform in the speech signal, it is necessary to add window processing after framing. The windowing of each short segment of the speech signal is equivalent to some kind of operation or transformation of each short segment, the specific formula is as follows:

$$Q_n = \sum_{m=-\infty}^{\infty} T[s(n)]\omega(n-m) \tag{5}$$

Where $T[\,]$ denotes some kind of transformation, linear or nonlinear can be, $s(n)$ is the input speech signal sequence, and Q_n is a time series obtained after all the segments have been processed.

The most commonly used window functions include Hamming window, rectangular window and Hanning window, which are defined respectively:

(1) Hamming window

$$\omega(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{N-1}\right) & 0\le n\le N-1\\ 0 & \text{other}\end{cases} \tag{6}$$

(2) Rectangular windows

$$\omega(n)=\begin{cases}1 & 0\le n\le N-1\\ 0 & \text{other}\end{cases} \tag{7}$$

(3) Hanning Window

$$\omega(n)=\begin{cases}0.5\left[1-\cos\left(\dfrac{2\pi n}{N-1}\right)\right] & 0\le n\le N-1\\ 0 & \text{other}\end{cases} \tag{8}$$

4) Endpoint Detection. Speech signal endpoint detection is a key link in the speech recognition process, the goal is to automatically detect the endpoints of the speech signal (i.e., the start and end points), and its algorithm will directly affect the performance of speech recognition. The start point and end point of the speech signal are accurately determined in order to store and process the useful speech signal. Experiments have shown that the short-time energy and short-time average over-zero rate dual-threshold endpoint detection method can more accurately detect the endpoints of valid speech, and is also a more commonly used and effective endpoint detection method. Dual-threshold comparison method with short-time energy E and short-time average zero crossing rate Z as a feature, effectively combining the advantages of E and Z, can accurately detect the endpoints of speech signals, not only can exclude the noise interference of the nine voice segments, but also reduce the processing time of the system to improve the real-time processing of the system, so as to improve the performance of the processing of speech signals. In this paper, this method is used for endpoint detection.

In the double threshold endpoint detection method, the characteristics of short-time energy E and short-time average over-zero rate Z are calculated as follows, respectively:

(1) The short-time energy E. The short-time energy of the speech signal $s(n)$ is defined as:

$$E_n=\sum_{m=-\infty}^{\infty}\left[s(n)\omega(n-m)\right]^2 \tag{9}$$

where $\omega(n)$ is the window function of the Hamming window.

For Eq. If we make $h(n)=\omega^2(n)$, we have

$$E_n=\sum_{m=-\infty}^{\infty}s^2(n)h(n-m)=s^2(n)*h(n) \tag{10}$$

From Eq. (10), the short-time energy of the window transform is equivalent to the output of the "speech-squared" signal through a linear filter with a unit sampling response of $h(n)$. The implementation process is as follows:

$$\xrightarrow{s(n)}(\cdot)^2\xrightarrow{s^2(n)}h(n)\xrightarrow{E_n} \tag{11}$$

For a particular frame of speech signal marked by n the short time average energy E_n is equation (12):

$$E_n=\sum_{m=n-N+1}^{n}[s(m)\omega(n-m)]^2 \tag{12}$$

(2) Short-time average zero crossing rate Z. The short-time average zero crossing rate is defined in equation (13):

$$Z_n=\sum_{m=-\infty}^{\infty}Sgn[s(m)]-Sgn[s(m-1)] \tag{13}$$

where Sgn[•] is the sign function, i.e., $Sgn[s(n)]=\begin{cases}1 & s(n)\ge 0\\ -1 & s(n)<0\end{cases}$, $s(n)$ are speech signals. Then:

$$Z_n = \sum_{m=-\infty}^{\infty} |Sgn[s(m)] - Sgn[s(m-1)]| \omega(n-m) = |Sgn[s(n)] - Sgn[s(n-1)]| * \omega(n) \quad (14)$$

where $\omega(n)$ is the window function.

Its realization flow is as follows:

$$\xrightarrow{s(n)} Sgn[\cdot] \rightarrow \text{First difference} \rightarrow |\cdot| \rightarrow \omega(n) \xrightarrow{Z_n} \quad (15)$$

The short time segment at the beginning of the speech signal is a uniformly distributed background noise signal. When using the double threshold comparison method for endpoint detection, it is necessary to calculate the over-zero rate threshold Z_cT and the high and low energy thresholds ETL (low-energy threshold) and ETU (high-energy threshold) based on the beginning of the "silent" segment as thresholds in order to realize the accurate detection of endpoints.

The over-zero rate threshold can be expressed as:

$$Z_cT = min(IF, Z_c + 2\sigma_{zc}) \quad (16)$$

Where IF is the empirical value, this paper takes $IF = 25$. Z_c, σ_{zx} is the mean and standard deviation of the zero crossing rate of the initial "silent" section.

For ETL (low-energy valve) and ETU (high-energy valve), need to calculate the "silent" section of the short-term average energy, the maximum energy value is recorded as $E_{\max}$, the minimum energy value is recorded as $E_{\min}$:

$$I1 = 0.03*(E_{\max} - E_{\min}) + E_{\min} \quad (17)$$

$$I2 = 4*E_{\min} \quad (18)$$

Then there is:

$$ETL = \min(I1, I2) \quad (19)$$

$$ETU = 5*ETL \quad (20)$$

When using Z_cT and ETL and ETU as thresholds for detection, let the starting frame be $N1$, then the energy E_{NI} and the over-zero rate Z_{NI} at frame $N1$ satisfy $ETU > E_{N1} > ETL$, $E_{NI+1} > ETU$, and $Z_{N1} > Z_cT$ at the same time.

The energy E_{N2} and the over-zero rate Z_{N2} at the end frame $N2$ satisfy both $E_{N2} < \frac{E}{T}L + ETUk$ (adjustment factor $k = 4$), $Z_{NI} < Z_cT$.

2.1.3. Speech feature parameter extraction. Speech signal characteristic parameter extraction is to remove the redundant information that is irrelevant to speech processing, and analyze and process the speech signal. Mel Frequency Cepstrum Coefficient (MFCC) simulates the response of the human ear to speech signals of different frequencies based on the mechanism of human hearing [39]. In fact, the sensitivity of the human ear's response to speech signals of different frequencies is different, similar to a special nonlinear system, which is basically logarithmic.

The extraction process of MFCC feature parameters is shown in Fig. 2.

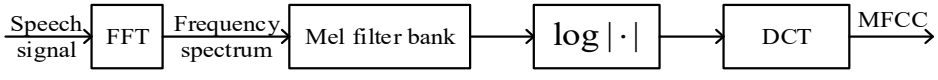

Fig. 2. MFCC feature parameter extraction process

The algorithm for MFCC speech feature parameter extraction is as follows.

1) Fast Fourier Transform (FFT) is shown in equation (21):

$$X[k]=\sum_{n=0}^{N-1}x[n]e^{-j\frac{2\pi}{N}nk},k=0,1,2,...,N-1 \tag{21}$$

Where $X[n],n=0,1,2,...,N-1$ is a discrete frame of speech sequence obtained after sampling and N is the frame length.

$X[k]$ is the complex sequence of N points, and then take the mode of $X[k]$ to get the signal amplitude spectrum $|X[k]|$.

2) Convert the actual frequency scale to Mel frequency scale as in equation (22):

$$Mel(f)=2597\lg(1+\frac{f}{700}) \tag{22}$$

Where, $Mel(f)$ is the Mel frequency and f is the actual frequency in Hz.

3) Configure the triangular filter bank and calculate the signal amplitude spectrum for each triangular filter $|X[k]|$ The filtered output is equation (23):

$$F(l)=\sum_{k=f_o(l)}^{f_h(l)}w_l(k)|X[k]|\quad l=1,2,...,L \tag{23}$$

Among them,

$$w_l(k)=\begin{cases}\dfrac{k-f_o(l)}{f_c(l)-f_o(l)}, & f_o(l)\le k\le f_c(l)\\ \dfrac{f_h(l)-k}{f_h(l)-f_c(l)}, & f_c(l)\le k\le f_b(l)\end{cases} \tag{24}$$

$$f_o(l)=\frac{o(l)}{\left[\frac{f_s}{N}\right]},f_h(l)=\frac{h(l)}{\left[\frac{f_s}{N}\right]},f_c(l)=\frac{c(l)}{\left[\frac{f_s}{N}\right]} \tag{25}$$

Where $w_l(k)$ is the filter coefficient of the corresponding filter, $o(l),c(l),h(l)$ is the lower limit frequency, center frequency and upper limit frequency of the corresponding filter on the real frequency axis, f_s is the sampling rate, L is the number of filters, and $F(l)$ is the filter output.

4) Do logarithmic operation on all filter outputs, and then further do the discrete cosine transform (DTC) to get the MFCC features, as in equation (26):

$$M(i)=\sqrt{\frac{2}{N}}\sum_{l=1}^{L}\log F(l)\cos[(l-\frac{1}{2})\frac{i\pi}{L}]\quad i=1,2,...,Q \tag{26}$$

where Q is the order of the MFCC parameter, in this paper we take 13, and $M(i)$ is the resulting MFCC parameter.

2.2. Speech recognition methods based on CTC and feature enhancement

2.2.1. Speech feature enhancement. Data quality and size are critical to the training results of the model, to improve data diversity, the work in this paper takes the approach of feature enhancement to expand the training data in the face of low resources. The feature enhancement work in this paper considers the MFCC spectrogram as the baseline for enhancement, and uses a combination of three methods, namely time warping, frequency masking, and time masking, to achieve the purpose of expanding the training data by modifying the MFCC spectrogram.

Taking a piece of speech data F19311001010259_read_l. wav in Exam_ASR as an example, in a Mel speech spectrogram, the horizontal axis is also called the time axis, the vertical axis is called the frequency axis, and the time distortion is the deformation of the sequence in the time direction. In the Mel speech spectrogram generated for a given period of speech of length τ, the so-called time distortion is the distortion to the left or to the right by random points on the horizontal line passing through the center of the spectrogram during a period of time $(W, \tau - W)$, and the distortion is a random number between the distances W W that are $(0, W)$, and W that are also called the time distortion factor.

Frequency masking means masking the data in a certain frequency range in the Mel spectrogram, which is implemented as shown below:

$$MFCC[f_0 : f_0 + f, :] = 0, \tag{27}$$

where f takes the value of $(0, F)$ within the frequency masking range $[f_0, f_0 + f)$, F the frequency masking factor, and f_0 takes the value of $[0, v - f)$, where v is the number of channels at the Mel frequency.

Similar to frequency masking, time masking is to mask out the data in the Mel spectrogram at a certain period of time, which is implemented as in Equation (28):

$$MFCC[:, t_0, t_0 + t] = 0 \tag{28}$$

2.2.2. Speech feature encoding and decoding. The CTC and feature enhancement based speech recognition method proposed in this chapter combines 2D convolutional neural network and GRU in the encoder part, using convolutional neural network to extract spatial features and recurrent neural network to extract temporal features. In the decoder part, CTC-based decoding is used.

1) Convolutional neural network. Convolutional neural network is a variant structure developed from multilayer perceptual machine (MLP), which is one of the representative algorithms of deep learning, and has a successful application in extracting visual features in the field of image processing. In terms of model construction, convolutional neural networks are accelerated by local connection and weight sharing, which reduces the number of parameters on the one hand and the complexity of the model on the other [40]. Because the concept of receptive field exists in the convolutional neural network, it can effectively encode local features, the size of the receptive field of a particular layer is the size of the region in which individual elements in the feature map output by this layer of the network are mapped back to the original input features, and in general, the deeper the layer of the network is, the larger the receptive field will be.

In the speech recognition method proposed in this chapter, 2D convolution is used to encode the feature-enhanced Mel speech spectrogram to learn the spatial features of the Mel speech spectrogram, which provides a strong support for the model in the learning of local information.

2) Recurrent Neural Networks. Recurrent neural network is also one of the representative algorithms of deep learning, because of the introduction of directed loops in the structure, which can deal with the temporal relationship between the sequential data, so it has a natural advantage in the field of speech recognition, and can effectively model the temporal information in the audio sequence. However, recurrent neural networks also have obvious drawbacks, in the training process gradient is easy to appear gradient disappearance or explosion, although cropping gradient can cope with gradient explosion, but can not solve the problem of gradient decay. In order to solve the gradient disappearance and gradient explosion in the recurrent neural network caused by the inability to establish a long-distance dependence problem, researchers have made improvements to the network structure, the more widely used structures are LSTM, GRU and so on.

Both LSTM and GRU structures memorize information with long-term dependencies by using gating units, which makes up for the shortcomings of traditional RNNs built on long-distance dependencies,

and three gates, input gate, output gate and forget gate, are introduced in LSTM to control the transmission of information by using the Sigmoid function. The recurrent neural network structure used in the speech recognition method proposed in this chapter is GRU, which is a variant of LSTM that retains only two gates, the reset gate and the update gate, where the update gate is a combination of the forgetting gate and the input gate from LSTM, the reset gate controls how much of the internal state from the previous moment needs to be stored, which is used to model short-term dependencies in the sequence, and the update gate controls whether the current moment in the internal state is a copy of the internal state in the previous moment, which is used to model the long-term dependencies in the sequence.

$\{f_0, f_1, \ldots, f_n\}$ is the feature vector encoded by the convolutional neural network, the hidden state in the forward process is calculated by the GRU unit in the Forward Layer, and the hidden state in the reverse process is calculated by the GRU unit in the Backward Layer, and finally the feature vector encoded by the GRU with the timing information of the forward and backward directions is obtained $\{m_0, m_1, \ldots, m_n\}$. In order to better utilize the timing information in the speech data, the proposed speech recognition model in this chapter uses a two-layer GRU structure, the specific calculation formula is as follows In order to better utilize the timing information in the speech data, the speech recognition model proposed in this chapter uses a two-layer GRU structure, and the specific calculation formula is as follows:

$$h_i = f(w_1 f_i + w_2 h_{i-1}) \tag{29}$$

$$h_i^{'} = f(w_3 f_i + w_4 h_{i+1}^{'}) \tag{30}$$

$$m_i = g(w_5 h_i + w_6 h_i^{'}) \tag{31}$$

where h_i denotes the hidden state in the forward computation process and $h_i^{'}$ denotes the hidden state in the reverse computation process, and the finally obtained encoded feature vector m_i contains both forward and reverse information, which helps to obtain more accurate prediction results in the subsequent decoding.

3) CTC-based decoding. The CTC algorithm is an algorithm that solves the problem of misalignment between input and output sequences, and its problem-solving ability is a natural fit for speech recognition tasks.The CTC algorithm, in solving the problem of misalignment between input and output sequences, is equivalent to creating an exact mapping between the encoding result $X=[x_1, x_2, \cdots, x_T]$ and the sequence label $Y=[y_1, y_2, \cdots, y_U]$. For a given X, plot the distribution of all outputs that could be an accurate mapping Y, and then maximize the probability of a correct output in the distribution, i.e., compute:

$$Y^* = \arg\max_{Y} P(Y \mid X) \tag{32}$$

The word 'hello' may be represented as '-hh-elll-l-oo-' or '-h-ee-l-l-oo-' and so on during the decoding process, and each representation is an output path at the time of output. When an output path, the final a posteriori probability is the sum of all the paths, CTC uses the idea of dynamic programming to merge the paths when solving the final a posteriori probability, and the formula of the a posteriori probability is expressed as:

$$P(Y \mid X) = \sum_{F(L)=Y} P(L \mid X) \tag{33}$$

where L is the path in the intermediate result and $F(L) = Y$ means the path whose output sequence is Y.

Since there is an independence assumption in the CTC algorithm and the output variables at different moments are independent of each other, the a posteriori probability of path L over X is expressed as:

$$P(L \mid X) = \prod_{t=1}^{T} z_{L_t}^t \tag{34}$$

where T is the total length of the coding vector, corresponding to the duration of the speech data, and $z_{L_t}^t$ denotes the character selected by path L at moment t. Combining the above two formulas can be obtained:

$$P(Y \mid X) = \sum_{F(L)=Y} \prod_{t=1}^{T} z_{L_t}^t \tag{35}$$

2.2.3. Loss function. This model uses a CTC-based decoding method in the decoding phase, so the loss function of the model is

$$L = L_{CTC} = \arg\max \sum_{F(L)=Y} \prod_{t=1}^{T} z_{L_t}^t \tag{36}$$

The training process allows the model to learn the optimal model parameters by minimizing the loss L of the model [41].

3. Exploration of a New Mode of English Listening Teaching in Colleges and Universities

So where is the new mode of listening teaching or the breakthrough in listening teaching? In the past to change the single repeated mechanical "listening" drill mode, in addition to listening at the same time for oral vocal exercises - read aloud. Reading aloud is a way of reading, and it is a form of vocal expression when you are very well prepared, even when you are able to recite in front of an audience. The significance of reading aloud for the English learner in terms of listening is that, through reading aloud, the English learner is not only outputting language, but also acquiring language as input, because he can hear his own speech and see the corresponding characters. Therefore, reading aloud combines the sound, shape and meaning of language to a large extent, which enables English learners to establish a good and effective unity between the characters, meaning and pronunciation of English through repeated practice in the multi-angle language input and output, so as to develop all the skills of English in a coordinated and comprehensive way. Of course, this includes listening skills. On the contrary, the disadvantage of relying on listening alone to improve listening skills is that at this point only the sounds and meanings of English are linked. In addition, if you only read silently, you are only connecting the form of English with its meaning. Therefore, there is a loss of balance between the various language systems and skills. Reading aloud can improve listening comprehension because when English learners practice reading aloud, it is the result of the simultaneous use of multiple senses, such as the eyes, ears, mouth, etc. Reading aloud can improve listening comprehension by strengthening the connection between the characters and meanings of the language and its pronunciation, and by repeatedly practicing so that the learners can know the semantics of the language as soon as they hear the pronunciation, thus improving the level of listening comprehension. On the other hand, listening to the language in order to improve listening comprehension will only result in the phenomenon of listening to what one understands and never understanding what one doesn't understand, which is the common problem of being able to read but not understanding what one reads. Even if there is no or not enough other forms of input, after a period of time the listening level will be stagnant or even regress. This is because the human language input and output systems

and their corresponding sensory systems are relatively independent channels. Generally speaking, the utilization of one channel does not affect the other. In other words, an increase in the utilization of one channel will not lead to an increase in the functioning of another channel.

In addition, reading aloud training, like any other language training, should emphasize the word perseverance. The principle of "no progress, no regression" is an indispensable rule in language learning and utilization; otherwise, language ability will deteriorate. Therefore, only through long-term and repeated reading aloud training can the different language channels be harmonized and developed, so that the phonological, morphological and ideological correspondences of language can be formed and consolidated.

Finally, when reading aloud training, firstly, we should choose materials with slightly lower difficulty than our own silent reading, because the multi-sensory use of reading aloud will interfere with semantic understanding to a certain extent, and if we choose the more difficult to understand language materials, we may read aloud for half a day but we don't know what we are reading, and then the reading aloud training is not likely to improve the listening level. Secondly, reading aloud training should follow the principle of gradual progress. First start with simple, easy-to-understand, short and concise materials, and gradually choose some relatively difficult and long materials for reading aloud training. The same applies to the training time. Thirdly, it is better to choose authentic original texts and contents of your own interest for read-aloud training. Because on the one hand, if the input material is not authentic or has errors, it will affect the whole language learning and utilization. On the other hand, only the material that interests you can be read aloud with good efficiency and effect, otherwise, you will only give up halfway and eventually give up. Fourthly, it is very important for readers to pronounce the language accurately. If one's pronunciation is inaccurate, the harder one trains for reading aloud, the faster one's listening level will drop. Because they are used to listening to their own substandard or incorrect pronunciation, when they hear the standard pronunciation, they will feel that they are listening to another foreign language. Therefore, in listening classroom teaching, we should not only use reading aloud training to improve students' English listening level, but also use reading aloud training appropriately, so that English learners' language skills can be developed in a comprehensive and balanced way.

The relationship between language skills is one of mutual influence and mutual constraint, so English learners should make efforts to improve the overall language level. Further, the training of language skills should also focus on the improvement of the overall language level, and should clearly realize the relationship between the skills, so as not to get into the misunderstanding of "treating the symptoms when they are not there".

4. Analysis of the Quality Improvement of English Listening Teaching Based on Speech Recognition Technology

4.1. Speech recognition technology performance validation analysis

4.1.1. Experimental setup:

1) Experimental setup. Since the experiments are to test the training of joint speech enhancement with speech recognition, the training set and the tests use a mixture of different signal-to-noise ratios in order to properly assess the robustness of the model. The experiments use AISHELL-1 as the pure speech dataset. Noise from DNS Challenge is used to get training set with noisy speech dataset by mixing in pure speech dataset with signal to noise ratios of -2.5dB, 2.5dB, 7.5dB, 12.5dB. A test set of the noise-containing speech dataset is obtained by mixing the noise in the pure speech dataset using the noise in the DNS Challenge at signal-to-noise ratios of -5dB, 0dB, 5dB, 10dB. Enhancement operations are performed on the noise-containing speech dataset and the pure speech dataset using the methods in this paper to obtain the enhanced pure speech dataset and the enhanced noise-containing speech dataset.

2) Experimental environment. In the noise robust speech recognition experiments with fused augmented features, an efficient computing environment is used to support the training process of the model in order to guarantee the training efficiency of the speech recognition model. The experimental environment setup for noise robust speech recognition with fusion-enhanced features is shown in Table 1. In terms of hardware facilities, a well-performing server with an Intel Xeon E3-1275 processor, which comes with a high frequency of 3.8 GHz, is coupled with a large memory capacity of 64 GB and an Nvidia Tesla V100 graphics processor equipped with 32 GB of video memory, which is a GPU that excels in terms of the computational performance of the graphic processing unit. In terms of software configuration, Ubuntu 16.04 was chosen as the base operating system, and the model construction and training were completed using the Python 3.9.10 programming language and the Py Torch 2.0.1 deep learning framework using the GPU-accelerated version.

Table 1. Experimental environment setting

CPU	Intel Xeon E3-1275 @ 3.8GHz
GPU	Nvidia Tesla V100 @ 32GB
Memory	64GB
Programming language	Python 3.9.10
Depth learning framework	PyTorch 2.0.1
Operating system	Ubuntu 16.04

3) Parameter setting. The FBank features of the speech signal are used as the back-end input in the noise robust speech recognition experiments incorporating enhanced features. Given that the perception mechanism of the human ear on the sound spectrum presents nonlinear characteristics, the FBank feature extraction technique simulates this auditory characteristic, thus more closely resembling the human auditory processing, which in turn helps to improve the overall accuracy of speech recognition. In the pre-processing stage of feature extraction, the time-domain signal is first converted to the frequency domain by STFT. Specifically, the parameters of the STFT in the experiment were configured to use a window length of 350 sampling points and a frame shift distance of 160 sampling points. In order to minimize the effect of spectral leakage, the Hanning window function was used for processing in this study. Subsequently, for each frame of the signal, its energy spectrum is computed and feature extraction is performed using a Mayer filter bank containing 135 filters, resulting in the FBank feature vector input to the speech recognition model. The signal processing parameter settings in the enhanced matched noise robust speech recognition experiments are shown in Table 2. Three different architectures are used as the back-end speech recognition models in the experiments, which are RNN-T, LAS and Conformer.The Conformer model is an optimized and extended model based on the Transformer model, in which the introduction of new structures such as depth-separable convolution, multi-branch convolution, and convolutional attention improves the performance and computational efficiency of the model.

Table 2. Enhanced matching noise noise robust voice recognition experiment

Parameter name	Set
Sampling rate	18000
The size of the STFT window	350
STFT frame shift	160
Window function	Hanning
MEL filter number	135

During the training process of the speech recognition model, the training parameters of the enhanced matched noise robust speech recognition are set as shown in Table 3. In order to update the model parameters, the Adam optimization algorithm was chosen and the exponential decay rate

parameter β_1 for first-order moment estimation was adjusted to 0.9 while the exponential decay rate parameter β_2 for second-order moment estimation was set to 0.9. 68 samples were used in each batch iteration as a way to ensure a balance between training efficiency and memory usage. The model was trained over 100 iteration cycles to ensure the adequacy of training. To efficiently adjust the weights, the learning rate was set at 1e-3. In addition, a gradient trimming threshold of 0.9 was set to prevent the gradient explosion problem. To avoid overfitting, a weight decay strategy of 5e-5 was adopted for the experiments. With these parameter configurations, the experiments ensure that the speech enhancement model can converge successfully in the training phase.

Table 3. Enhanced matching noise and sound speech recognition training parameter Settings

Parameter name	set
Optimizer	Adam
Adam optimizer β_1	0.9
Adam optimizer β_2	0.9
Batch Size	68
Epoch	100
Learning rate	1e-3
Gradient cutting threshold	0.9
Weight attenuation coefficient	5e-5

4.1.2. Analysis of experimental results:

1) Fusion Enhanced Feature Experiment. The effects of co-training and fusing the networks at signal-to-noise ratios of -5dB, 0dB, 5dB and 10dB on the performance are shown in Tables 4 to 7. The weight μ for the speech recognition task is set to 0.5, as this parameter configuration enables the models to reach convergence in the experiments, and the value of μ will be further explored in subsequent experiments.

By comparing Experimental Group I with Experimental Group II, it can be found that the addition of the fusion network has led to the improvement of the performance of the RNN-T, LAS and Conformer models under different SNR conditions. In particular, at a SNR of -5 dB, the addition of the fusion network resulted in a 4.57% reduction in the CER of the RNN-T model, a 3.73% reduction in the CER of the LAS model, and a 4.19% reduction in the CER of the Conformer model. This shows that using the augmentation network to connect the front and back ends brought performance improvement to all models, providing the models with features that are more suitable for the speech recognition task Comparing Experimental Group I with Experimental Group III, it can be found that performing joint training under different signal-to-noise ratio conditions resulted in improved performance for the RNN-T, LAS, and Conformer models. In particular, at a signal-to-noise ratio of -5 dB, the addition of the fusion network resulted in a 1.62% reduction in the CER of the RNN-T model, a 2.02% reduction in the CER of the LAS model, and a 1.78% reduction in the CER of the Conformer model. This indicates that performing joint training resulted in performance gains for all models. However, by comparing Experimental Group II with Experimental Group III, it can be seen that the improvement brought by joint training is not as large as the improvement brought by adding the fusion network. By comparing experimental group IV with the other experimental groups, it can be found that under different SNR conditions, performing joint training along with the addition of the fusion network resulted in the best performance of the RNN-T, LAS and Conformer models. In particular, at a SNR of -5dB, the addition of the fusion network resulted in a 5.69% reduction in the CER of the RNN-T model, a 6% reduction in the CER of the LAS model, and a 5.92% reduction in the CER of the Conformer model. This shows that the addition of the fusion network and joint training leads to the best performance for the model.

The fusion network combines the enhanced signal information with the original signal information to provide better input features to the recognition model. Co-training solves the sub-optimization

problem by simultaneously training the speech enhancement and speech recognition models end-to-end. Both fusion network and joint training can reduce the character error rate of the speech recognition model under different signal-to-noise ratios thus improving the robustness of the model. The fusion network outperforms the co-training when used alone, but the best results are achieved when the two are used together.

Table 4. The effect of the -5db signal-to-noise ratio on speech recognition

Experimental group	Joint training	Fusion network	CER(%)		
			RNN-T	LAS	Conformer
1	✗	✓	47.16	46.37	44.2
2	✗	✓	42.59	42.64	40.01
3	✓	✗	45.54	44.35	42.42
4	✓	✗	41.47	40.37	38.28

Table 5. The effect of the 0db signal-to-noise ratio on speech recognition

Experimental group	Joint training	Fusion network	CER(%)		
			RNN-T	LAS	Conformer
1	✗	✓	38.25	36.98	36.27
2	✗	✓	32.25	33.09	30.67
3	✓	✗	33.81	35.19	32
4	✗	✓	30.67	33.03	29

Table 6. The effect of the 5db signal-to-noise ratio on speech recognition

Experimental group	Joint training	Fusion network	CER(%)		
			RNN-T	LAS	Conformer
1	✓	✗	35.32	34.63	32.39
2	✓	✓	28.14	29.95	24.94
3	✓	✗	32.95	31.76	28.62
4	✓	✓	25.08	28.68	21.61

Table 7. The effect of the 10db signal-to-noise ratio on speech recognition

Experimental group	Joint training	Fusion network	CER(%)		
			RNN-T	LAS	Conformer
1	✗	✗	28.28	25.94	23.14
2	✗	✓	23.41	20.14	19.43
3	✓	✗	25.21	23.2	20.69
4	✗	✓	20.17	17.96	15.04

2) Speech Recognition Task Weighting Experiment. This experiment used the previously best performing configuration, i.e., the inclusion of the fusion network and joint training, in order to explore appropriate speech recognition task weight μ fetching values. Given that a comprehensive search for all weighting values is a daunting task, tests were conducted by taking 0.1 steps over a range of values from 0.3 to 0.8. The results were obtained by averaging the results obtained from the test set at -5dB, 0dB, 5dB, and 10dB SNR. The effect of speech recognition weights values on the model performance is shown in Table 8. When μ is set to 0.6, the RNN-T and LAS models exhibit optimal CERs of 36.94% and 35.59%, respectively. When λ is 0.5, the Conformer model achieves the best CER of 34.86%. When λ increases to 0.7 or decreases to 0.4, all models show a downward trend in performance. When λ increases to 0.8 or decreases to 0.3, all models get the worst performance.

This is perhaps an imbalance in the weighting of the training tasks interfering with the overall performance of the main speech recognition system. This leads to the conclusion that the speech recognition task weights μ are best taken as 0.5 or 0.6.

Table 8. The effect of speech recognition weight on model performance

μ	CER(%)		
	RNN-T	LAS	Conformer
0.3	41.79	42.53	40.7
0.4	37.95	37.3	36.82
0.5	37.14	36.22	34.86
0.6	36.94	35.59	35.29
0.7	39.88	39.47	35.58
0.8	42.95	41.24	37.1

4.2. Speech signal visualization

In addition to the literal content of the language, verbal communication will also contain some phonetic details, the main one of which is the emotion of the speaker. The same content expressed through different emotions, the meaning of the statement may be the opposite. For example, this sentence "even if it rains, go", in the mood of anger, it can be expressed in order to reproach the meaning. In a happy mood, it can convey a feeling of anticipation and joy. In the mood of surprise, one can express a question. And in the mood of fear, it can contain the idea of not wanting to perform. So, although the content of a sentence is the same, the meaning expressed in different contexts can be completely different. Based on comparing the speech signals of the same sentence under different emotions, the emotions contained in the speech signals are analyzed. In this paper, the fundamental frequency of the speech signal is used as a feature to analyze the emotion contained in the speech. For example, in calm and sad moods, the fluctuation of speech will be relatively smooth. In contrast, in happy and angry moods, the fluctuations of speech will be significantly stronger and in a higher range of fundamental frequency. Fear fluctuations and fundamental frequency ranges are in between the first two categories. Speech in the emotion of anger is shown in Figure 3. It can be seen that the fundamental frequency can be as high as 450 Hz and is overall above 100 Hz with a strong degree of fluctuation. The speech under the happy emotion is shown in Figure 4. The overall frequency of the base tones are higher and the fluctuation is obvious.

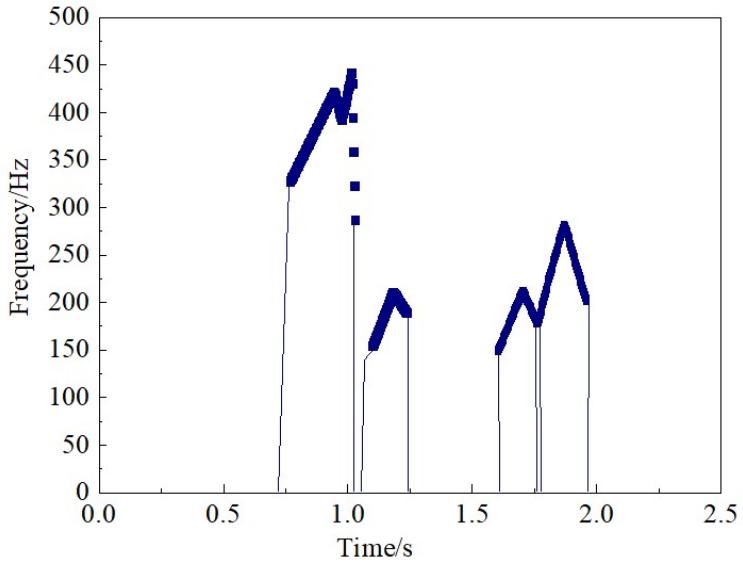

Fig. 3. Voice of anger

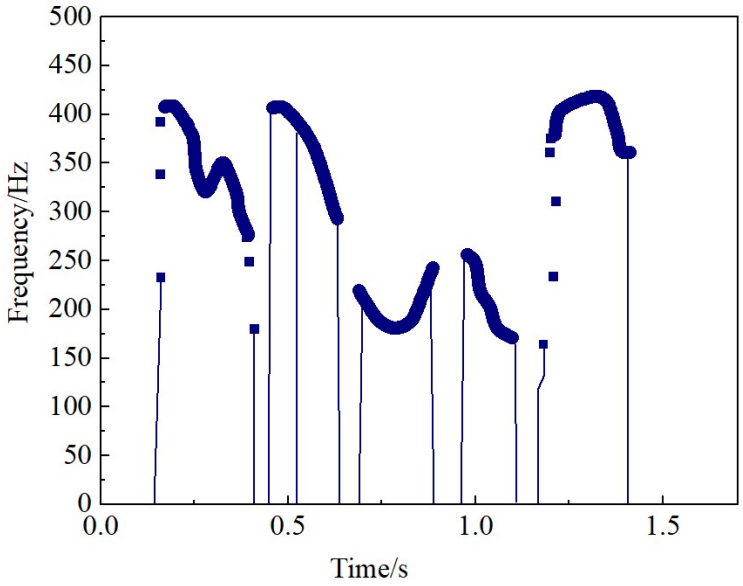

Fig. 4. Voice of happiness

Sad emotions and calm voices are shown in Figure 5 and Figure 6. Except at the beginning, the overall speech fundamental frequency was below 250 Hz and varied smoothly. By comparison with the anger and happy emotions, it is clear that the basal tone frequency is lower and less volatile.

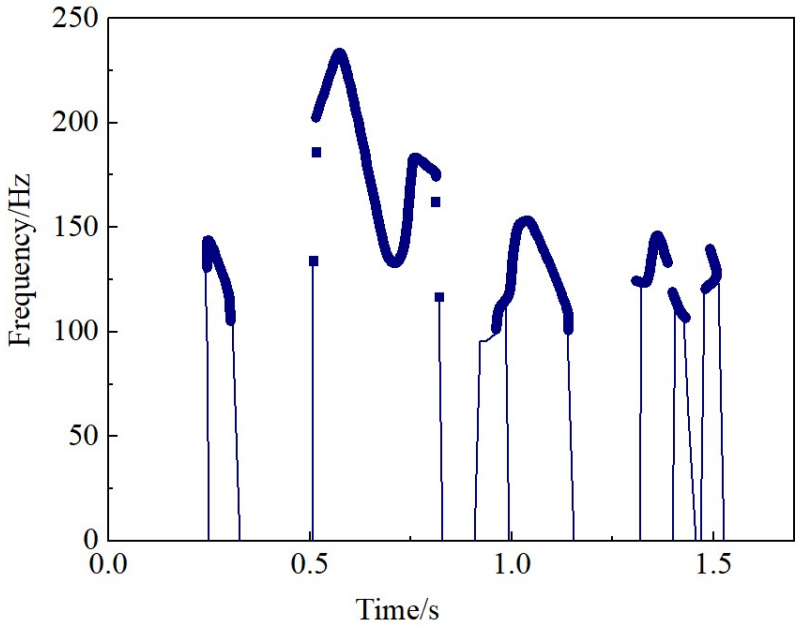

Fig. 5. Voice in sad mood

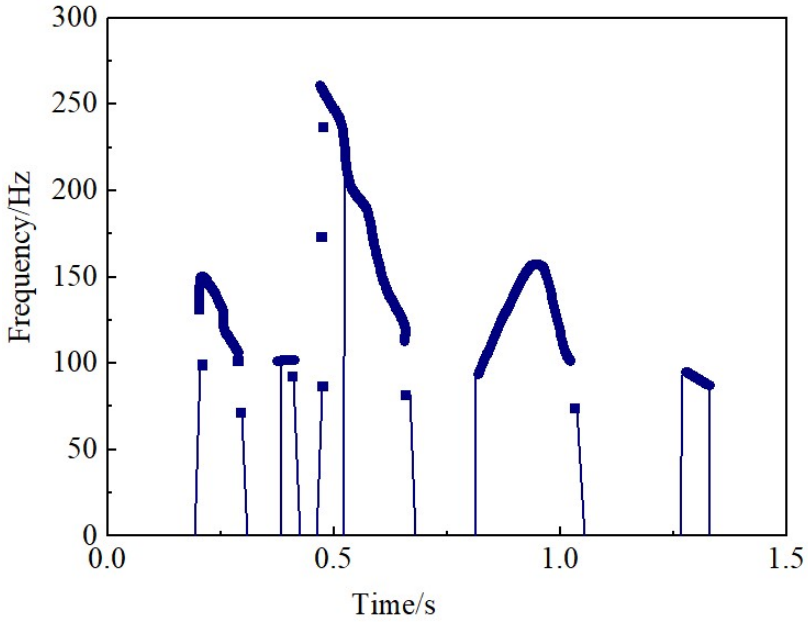

Fig. 6. Quiet voice

4.3. Effectiveness of Teaching Quality Improvement

4.3.1. Experimental Methods and Procedures. The experimental methodology mainly includes the experimental research objectives, research hypotheses, research subjects and materials used for testing. The research objectives specifically include two points: firstly, to test whether there is a significant difference between the grades under the teaching mode proposed in this paper and the grades under the traditional teaching mode, and secondly, to test the degree of correlation between the grades under the teaching mode proposed in this paper and the grades under the traditional teaching mode. According to the research objectives, the research hypotheses also include two points: hypothesis one is that there is no effect of the two teaching modes on the performance of the candidates, i.e., the performance of the two groups is equal in general and there is no significant difference. Hypothesis two is that there is no correlation between the grades in the teaching mode proposed in this paper and the grades in the traditional teaching mode. The research object is 70 foreign students in a university, in order to ensure the accuracy and effective implementation of the experiment, before formally carrying out the experiment according to the results of the most recent examination, the 70 students were divided into two groups of equal level, that is, the teaching mode group proposed in this paper and the traditional group, 35 students in each of the two groups, and each group of men and women in equal proportions, based on which, the students of each group were further ranked according to their scores from highest to lowest. The students in each group were further ranked in descending order of performance, and the numbers 1 through 35 were used sequentially as their school numbers used in the test. The materials used for the test include the materials in the teaching mode proposed in this paper and the online test tool developed in this experiment. These two materials (or tools) are only presented in different forms, but their question types and contents are exactly the same, consisting of four modules of listening, speaking, reading, and writing, and each module has six questions with one point for each question. After the preparation of the experimental materials, the two groups of students were organized to conduct the teaching mode proposed in this paper and the traditional teaching mode in the specified time, and after the test, some of the students were interviewed and asked about their attitudes and views on speech synthesis technology and speech recognition technology, and finally, the data collected from the test were sorted

out accordingly and analyzed by the SPSS statistical software for social sciences in order to verify the hypotheses.

4.3.2. Analysis of experimental results:

1) Descriptive statistics. The descriptive statistics of the results of the teaching model group and the traditional group proposed in this paper are shown in Table 9. From this information in the table, we draw several conclusions: firstly, for the two groups of candidates, the mean of the total performance of the candidates in the teaching model group proposed in this paper (17.703) is higher than the mean of the total performance of the candidates in the traditional group (17.443), and the mean of the listening performance (4.9), the mean of the reading performance (4.963), and the mean of the writing performance (4.07) in the teaching model group proposed in this paper are higher than the mean of the listening performance (4.8) and the mean of the reading performance (4.963) in the traditional group, respectively.) were higher than the mean values of listening (4.8), reading (4.923) and writing (3.59) scores of the candidates in the traditional group, respectively; on the contrary, the mean value of speaking score of the candidates in the group with the teaching model proposed in this paper (4.26) was lower than the mean value of speaking score of the candidates in the traditional group (4.18). Secondly, for the two groups of candidates, the standard deviation (3.793), standard deviation (1.42), standard deviation (0.964) and standard deviation (1.808) of the total scores of the test takers in the teaching mode group were higher than those of the test takers in the traditional group (2.625) and the standard deviation of the speaking score (1.012), respectively. The standard deviation of reading scores (1.185) and writing scores (1.863) indicate that the total scores, speaking scores, reading scores and writing scores of the test takers in the teaching mode proposed in this paper are more discrete. On the contrary, the standard deviation of the listening scores of the candidates in the teaching model group proposed in this paper (1.091) was lower than that of the listening scores of the candidates in the traditional group (0.931), which suggests that the listening scores of the candidates in the traditional group had a greater degree of dispersion. Finally, this study analyzed the skewness and kurtosis of the total scores and the scores of each part of listening, speaking, reading and writing of the teaching mode group and the traditional group proposed in this paper through SPSS, and the purpose of analyzing the skewness and kurtosis is to test whether the data obey normal distribution, and it is important to note that there are many ways to determine whether the data obey normal distribution, and the reason why this study chose to use the skewness and kurtosis is that the sample size of this experiment is N 35, whose value is less than 100.Regarding skewness and kurtosis, specifically, skewness refers to the degree that describes the deviation of the data distribution from symmetry, skewness = 0 when the distribution of the sample data conforms to normal distribution, and skewness > 0 or skewness < 0 when the distribution of the sample data does not conform to normal distribution.Kurtosis refers to the degree of steepness and slowness that describes the pattern of the distribution of the data, kurtosis = 0 when the distribution of the sample data conforms to normal distribution, and kurtosis = 0 when the distribution of the sample data does not conform to normal distribution. When the distribution of the sample data does not conform to normal distribution, kurtosis>0 or kurtosis<0. The test of normal distribution using skewness and kurtosis can be realized by calculating its corresponding Z-value, and this paper has calculated all the skewness data and kurtosis data in the table, and finally found that the total scores and the scores of each part of the listening, speaking, reading, and writing of the teaching mode group and the traditional group proposed in this paper obeyed the normal distribution, which was The next independent sample t-test and correlation analysis laid the foundation.

Table 9. Descriptive statistics of the composition of paper and pen and machine

	N	Mean	Standard Deviation	Degree Of Bias	Kurtosis

	Statistic	Statistic	Statistic	Deviation Value	Standard Error	Deviation Value	Standard Error
Machine Score	35	17.443	2.625	-0.748	0.416	0.507	0.842
Test Performance	35	4.8	0.931	-0.378	0.416	-0.139	0.842
Test Performance	35	4.18	1.012	-0.543	0.416	0.128	0.842
Test Performance	35	4.923	1.185	-0.283	0.416	-0.61	0.842
Machine Test Writing	35	3.59	1.863	-0.585	0.416	-0.888	0.842
Paper Record	35	17.703	3.793	-0.761	0.416	0.389	0.842
Paper Note	35	4.9	1.091	-0.45	0.416	0.106	0.842
Oral Performance Of Paper	35	4.26	1.42	-0.597	0.416	-0.551	0.842
Paper Writing	35	4.963	0.964	-0.957	0.416	-0.887	0.842
Paper Writing	35	4.07	1.808	-0.419	0.416	-0.657	0.842

2) Comparability analysis of the test scores of the teaching model group and the traditional group proposed in this paper. On the basis of descriptive statistical analysis, in order to verify hypothesis one, this study further conducted independent samples t-tests on the total scores of the two groups of test takers and the scores of each part of listening, speaking, reading, and writing, respectively, in order to test whether there is a significant difference between the scores of the groups, and the comparison of the total scores of the teaching model group and the traditional group proposed in this paper is shown in Table 10. In the table, the Sig. value (0.069) of the "Levene test of the variance equation" is greater than 0.05, indicating that the overall variance of the two sets of results is equal, and then the results of using the first row labeled "assuming equal variance" are determined, and on this basis, the "t-test of the mean equation" is further looked at The value of 0.775 is much greater than 0.05, so it can be concluded that although the mean of the traditional group (17.443) is lower than the mean of the proposed teaching mode group (17.703), there is no significant difference between the two groups.

Table 10. Comparison of paper and pen and machine team

		Levene of the variance equation		T test of the mean equation				
		F	Sig.	Sig.(Double side)	Mean difference	Standard error value	95% confidence interval of the difference	
							Lower limit	Upper limit
Total score	Let's say the variance is equal	3.226	0.069	0.775	-0.265	0.900	-2.058	1.540
	Let's say that the variance is not equal			0.775	-0.265	0.900	-2.061	1.548

The comparison of the scores of the listening, reading and writing components in the teaching model group and the traditional group proposed in this paper is shown in Table 11. On the basis of descriptive statistics, the Sig. (bilaterally) value corresponding to the listening achievement is 0.685, which is much larger than 0.05, so it can be concluded that although the mean value of the listening achievement of the traditional group (4.8) is lower than that of the listening achievement of the teaching model group proposed in this paper (4.9), there is no significant difference between the two groups' achievements. The Sig. (two-sided) value corresponding to the speaking performance is 0.577, which is much larger than 0.05, so it can be concluded that although the mean value of the speaking performance of the traditional group (4.18) is higher than that of the speaking performance of the teaching model group proposed in this paper (4.26), there is no significant difference between the performances of the two groups. The Sig. (two-sided) value corresponding to the reading performance is 0.435, which is much larger than 0.05, so it can be concluded that although the mean value of the reading performance of the traditional group (4.923) is lower than the mean value of the reading performance of the teaching model group proposed in this paper (4.963), there is no significant difference between the two groups. The Sig. (two-sided) value corresponding to the writing achievement is 0.715, which is much greater than 0.05, so it can be concluded that although the mean value of the writing achievement of the traditional group (3.59) is lower than the mean value of the writing achievement of the instructional model group proposed in this paper (4.07), there is no significant difference between the two groups' achievements.

Table 11. The paper pen and the machine test group are compared to the performance

		Levene of the variance equation		T test of the mean equation				
		F	Sig.	Sig.(Double side)	Mean difference	Standard error value	95% confidence interval of the difference	
							Lower limit	Upper limit
Hearing score	Let's say the variance is equal	2.765	0.105	0.685	-0.1	0.241	-0.595	0.395
	Let's say that the variance is not equal			0.685	-0.1	0.241	-0.597	0.397
Oral performance	Let's say the variance is equal	1.856	0.168	0.577	-0.2	0.355	-0.502	0.901
	Let's say that the variance is not equal			0.577	-0.2	0.355	-0.505	0.902
Reading score	Let's say the variance is equal	0.036	0.865	0.435	-0.2	0.269	-0.726	0.325
	Let's say that the variance is not equal			0.435	-0.2	0.269	-0.726	0.325
Writing score	Let's say the variance is equal	2.593	0.121	0.715	-0.165	0.435	-1.021	0.705
	Let's say that the variance is not equal			0.715	-0.165	0.435	-1.023	0.706

3) Correlation analysis of the test scores of the teaching model group and the traditional group proposed in this paper. On the basis of the above analysis, in order to test hypothesis two, this study further conducted correlation analysis on the total scores and the scores of each part of listening, speaking, reading and writing of the test takers of the two groups respectively, in order to test whether there is any correlation between the scores of each group, and the correlation between the total scores of the teaching-mode group and the traditional group proposed in this paper and the scores of each part of listening, speaking, reading and writing is shown in Table 12 (** denotes a significant

correlation at the level of 0.01, and * indicates significant correlation at 0.05 level). The following conclusions can be drawn from the table: firstly, the significance between the total scores and the scores of each part of listening, speaking, reading and writing of the teaching model group and traditional group proposed in this paper are <0.05, so it can be considered that there is a significant correlation between the total scores and the scores of each part of listening, speaking, reading and writing of the teaching model group and traditional group proposed in this paper. Secondly, analyzing the Pearson's correlation coefficient on the basis of the significance analysis, it was found that the correlation coefficient between the total scores of the paper-and-pencil group and the traditional group was 0.662, and the absolute value of the correlation coefficient was between 0.6 and 0.8, which indicated that there was a strong correlation between the total scores of the two modes, and the correlation coefficient of the scores of the teaching-mode group of the teaching-mode group of the paper proposed in the present paper and the scores of the traditional group of the teaching-mode group in the parts of listening, speaking, reading, writing were all within 0.5 to 0.5. The absolute values of the correlation coefficients are all between 0.4 and 0.6, indicating that there is a moderate correlation between the grades of the listening, speaking, reading and writing components in these two modes.

The synthesis of the above statistics and analysis of the results shows that although there is a difference between the test scores of the teaching mode group proposed in this paper and the traditional group, it is not statistically significant. At the same time, there is a strong or moderate correlation between the scores of the teaching mode group proposed in this paper and those of the traditional group, which indicates that the scores of most of the students did not change greatly due to the change in the testing mode, and thus the final conclusion can be drawn that the testing results of the teaching mode proposed in this paper are equivalent to those of the traditional teaching mode.

Table 12. Correlation of grades

	N	Pearson correlation coefficient	significance
The total results of the machine and the total results of paper	35	0.662**	0.000
Test performance of machine and paper pen hearing	35	0.405*	0.025
Test performance of machine and paper oral performance	35	0.451*	0.016
Test performance & paper writing	35	0.522**	0.002
Test writing grades & paper writing results	35	0.45*	0.016

5. Conclusion

In this study, a speech recognition technology research was carried out and performance testing experiments and empirical tests were conducted around the issue of improving the quality of listening instruction in English language teaching.

In the signal-to-noise ratio experiment at -5dB, it is found that the addition of the fusion network reduces the CER of the RNN-T model by 5.69%, while the LAS model reduces the CER by 6%, and the Conformer model reduces the CER by 5.92%. This shows that the model in this paper has the best performance compared to the other models.

By conducting an equivalence study between the proposed teaching model and the online system test on 70 international students in a certain university, the experimental results show that there is no significant difference between the total test scores and the listening, reading and writing parts of the proposed teaching model group and the traditional group, and the correlation between the total test scores and the listening, reading and writing parts of the traditional group and the teaching model group of the proposed teaching model group is a strong correlation or a moderate correlation. The correlation between the total scores of the test and the scores of each part of listening, reading and writing of the traditional group is strong or moderate. That is to say, the test results of the teaching model proposed in this paper are equivalent to those of the traditional teaching model. This proves that the overall quality of English teaching has been improved after applying the method of this paper.

References

[1] Krivosheyeva, G., Zuparova, S., & Shodiyeva, N. (2020). Interactive way to further improve teaching listening skills. Academic Research in Educational Sciences, (3), 520-525.

[2] Abdulrahman, T., Basalama, N., & Widodo, M. R. (2018). The Impact of Podcasts on EFL Students' Listening Comprehension. International Journal of Language Education, 2(2), 23-33.

[3] Chen, J. (2022). Reform of English writing teaching method under the background of big data and artificial intelligence. International Journal of e-Collaboration (IJeC), 19(4), 1-16.

[4] Yao, Y., & Ma, C. (2023). RETRACTED: A multimedia network English listening teaching model based on confidence learning algorithm of speech recognition. International Journal of Electrical Engineering & Education, 60(1_suppl), 3688-3701.

[5] Listiyaningsih, T. (2017). The influence of listening English song to improve listening skill in listening class. Academica: Journal of Multidisciplinary Studies, 1(1), 35-49.

[6] Yanhua, Z. (2020, June). The application of artificial intelligence in foreign language teaching. In 2020 International Conference on Artificial Intelligence and Education (ICAIE) (pp. 40-42). IEEE.

[7] Zhang, X., Peng, Y., & Xu, X. (2019, September). An overview of speech recognition technology. In 2019 4th International Conference on Control, Robotics and Cybernetics (CRC) (pp. 81-85). IEEE.

[8] Bhatt, S., Jain, A., & Dev, A. (2021). Continuous speech recognition technologies—a review. In Recent Developments in Acoustics: Select Proceedings of the 46th National Symposium on Acoustics (pp. 85-94). Springer Singapore.

[9] Kumar, T., Rajest, S. S., Villalba-Condori, K. O., Arias-Chavez, D., Rajesh, K., & Chakravarthi, M. K. (2022). An evaluation on speech recognition technology based on machine learning. Webology, 19(1), 646-663.

[10] Wang, D., Wang, X., & Lv, S. (2019). An overview of end-to-end automatic speech recognition. Symmetry, 11(8), 1018.

[11] Oh, E. Y., & Song, D. (2021). Developmental research on an interactive application for language speaking practice using speech recognition technology. Educational Technology Research and Development, 69(2), 861-884.

[12] Kim, N. Y. (2019). A study on the use of artificial intelligence chatbots for improving English grammar skills. Journal of Digital Convergence, 17(8).

[13] Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2024). 'Look, I can speak correctly': learning vocabulary and pronunciation through websites equipped with automatic speech recognition technology. Computer Assisted Language Learning, 37(5-6), 1335-1363.

[14] Huang, X. (2021, November). RETRACTED: The Design of College English Teaching Platform Based on

Artificial Intelligence. In Journal of Physics: Conference Series (Vol. 2066, No. 1, p. 012084). IOP Publishing.

[15] Liu, X., Xu, M., Li, M., Han, M., Chen, Z., Mo, Y., ... & Liu, M. (2019). Improving English pronunciation via automatic speech recognition technology. International Journal of Innovation and Learning, 25(2), 126-140.

[16] Manire, E., Kilag, O. K., Cordova Jr, N., Tan, S. J., Poligrates, J., & Omaña, E. (2023). Artificial Intelligence and English Language Learning: A Systematic Review. Excellencia: International Multi-disciplinary Journal of Education (2994-9521), 1(5), 485-497.

[17] Ren, H., Mao, X., Ma, W., Wang, J., & Wang, L. (2020). An English-Chinese machine translation and evaluation method for geographical names. ISPRS International Journal of Geo-Information, 9(3), 139.

[18] Liang, J., & Du, M. (2022). Two‐Way Neural Network Chinese‐English Machine Translation Model Fused with Attention Mechanism. Scientific Programming, 2022(1), 1270700.

[19] Takakusagi, Y., Oike, T., Shirai, K., Sato, H., Kano, K., Shima, S., ... & Katoh, H. (2021). Validation of the reliability of machine translation for a medical article from Japanese to English using DeepL translator. Cureus, 13(9).

[20] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., ... & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the national academy of sciences, 117(14), 7684-7689.

[21] Sahu, P., Dua, M., & Kumar, A. (2018). Challenges and issues in adopting speech recognition. Speech and Language Processing for Human-Machine Communications: Proceedings of CSI 2015, 209-215.

[22] 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.

[23] Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. Procedia Computer Science, 128, 32-37.

[24] Delić, V., Perić, Z., Sečujski, M., Jakovljević, N., Nikolić, J., Mišković, D., ... & Delić, T. (2019). Speech technology progress based on new machine learning paradigm. Computational intelligence and neuroscience, 2019(1), 4368036.

[25] Agustin, R. W., & Ayu, M. (2021). The impact of using Instagram for increasing vocabulary and listening skill. Journal of English Language Teaching and Learning, 2(1), 1-7.

[26] Rintaningrum, R. (2018). Investigating reasons why listening in English is difficult: voice from foreign. Asian EFL Journal, 20(11), 6-15.

[27] Suwannasit, W. (2019). EFL LEARNERS'LISTENING PROBLEMS, PRINCIPLES OF TEACHING LISTENING AND IMPLICATIONS FOR LISTENING INSTRUCTION. Journal of Education Naresuan University, 21(1), 345-359.

[28] Zhao, X., Wang, Y., Liu, Y., Xu, Y., Meng, Y., & Guo, L. (2019). Multimedia based Teaching Platform for English Listening in Universities. International Journal of Emerging Technologies in Learning, 14(4).

[29] Afriyuninda, E., & Oktaviani, L. (2021). THE USE OF ENGLISH SONGS TO IMPROVE ENGLISH STUDENTS'LISTENING SKILLS. Journal of English Language Teaching and Learning, 2(2), 80-85.

[30] Jiang, Q. (2018, June). The practice of English listening comprehension teaching. In 2018 3rd International Conference on Humanities Science, Management and Education Technology (HSMET 2018) (pp. 543-546). Atlantis Press.

[31] Atmowardoyo, H., & Sakkir, G. (2022). Online-based English Listening Skill Learning Model. Celebes Journal of Language Studies, 239-246.

[32] Graham, S. (2017). Research into practice: Listening strategies in an instructed classroom setting.

Language teaching, 50(1), 107-119.

[33] Yu, C. (2024, June). Application of Artificial Intelligence and Speech Recognition Technology in English Listening and Speaking Ability Assessment System. In 2024 2nd International Conference on Mechatronics, IoT and Industrial Informatics (ICMIII) (pp. 348-352). IEEE.

[34] Wei, L. (2019). Study on the application of cloud computing and speech recognition technology in English teaching. Cluster Computing, 22(Suppl 4), 9241-9249.

[35] Chen, X. (2024). Using Speech Recognition Algorithms to Improve Listening Training in College English Listening Instruction. Journal of Electrical Systems, 20(6s), 1775-1785.

[36] Lai, K. W. K., & Chen, H. J. H. (2024). An exploratory study on the accuracy of three speech recognition software programs for young Taiwanese EFL learners. Interactive Learning Environments, 32(5), 1582-1596.

[37] Geng, L. (2021). Evaluation model of college english multimedia teaching effect based on deep convolutional neural networks. Mobile Information Systems, 2021(1), 1874584.

[38] Myagmarsuren Orosoo,Namjildagva Raash,Mark Treve,Hassan Fareed M. Lahza,Nizal Alshammry,Janjhyam Venkata Naga Ramesh & Manikandan Rengarajan. (2025). Transforming English language learning: Advanced speech recognition with MLP-LSTM for personalized education. Alexandria Engineering Journal21-32.

[39] Congmin Mao & Sujing Liu. (2024). A Study on Speech Recognition by a Neural Network Based on English Speech Feature Parameters:Regular Papers. Journal of Advanced Computational Intelligence and Intelligent Informatics(3),679-684.

[40] Seong Gyun Leem,Daniel Fulford,Jukka Pekka Onnela,David Gard & Carlos Busso. (2024). Selective Acoustic Feature Enhancement for Speech Emotion Recognition With Noisy Speech.. IEEE/ACM transactions on audio, speech, and language processing917-929.

[41] Ji-Won Cho & Hyung-Min Park. (2016). Independent vector analysis followed by HMM-based feature enhancement for robust speech recognition. Signal Processing200-208.