Hindawi
Mobile Information Systems
Volume 2022, Article ID 6131572, 12 pages
https://doi.org/10.1155/2022/6131572

Research Article

A Machine Learning Assessment System for Spoken English Based on Linear Predictive Coding

Lu Wang

School of Liberal Education, Chengdu Jincheng College, Sichuan, Chengdu 611731, China

Correspondence should be addressed to Lu Wang; wanglu@scujcc.edu.cn

Received 3 August 2022; Revised 18 August 2022; Accepted 30 August 2022; Published 20 September 2022

Academic Editor: R. Mo

In the teaching of English, there is an increasing focus on practical communication skills. As a result, the speaking test component has received more and more attention from education experts. With the rapid development of modern computer technology and network technology, the use of computers to assess the quality of spoken English has become a hot topic of research in related fields at present. A machine learning assessment system based on linear predictive coding is proposed in order to achieve automatic scoring of spoken English tests. First, the principle of linear predictive coding and decoding is analyzed, and the traditional linear predictive coding and decoding algorithm is improved by using hybrid excitation instead of the traditional binary excitation. Second, the overall structure of the machine learning assessment system is designed, which mainly includes division into four modules: acoustic model acquisition module, speech recognition module, standard pronunciation transcription module, and decision module. Then, the speech recognition module is implemented by an improved linear predictive speech coding method to acquire the feature parameters of the speech signal and generate the speech feature vector. Finally, the convolutional neural network algorithm is used to train the speech features so as to implement the acoustic model acquisition module. The experimental results show that the improved linear predictive speech coding method yields more natural and higher intelligibility speech signals. The designed machine learning evaluation system is able to accurately detect information about the quality of the learner's pronunciation.

1. Introduction

The focus of modern English language teaching is on the development of students' general application skills, including listening and reading skills. Among these, speaking training and speaking assessment have received increasing attention. There are generally two types of assessment for speaking tests: an automated assessment and a manual assessment by experts. With the continuous development of random computer technology, automated assessment of speaking tests is beginning to be used in a variety of industries [1–6]. For example, speaking assessment systems can be used during telephone interviews to automatically score the English proficiency of interviewees. In addition, online teaching application scenarios in the education industry can use speaking assessment systems to automate the scoring of students' speaking quality. Automated speaking assessment systems can give objective scores based on the test taker's performance in a timely manner and are not subjectively influenced by personal factors [7, 8].

As competition in business continues to intensify, there is an increasing demand for complex talents. Companies require these people to have not only solid professional knowledge, but also to be able to express themselves proficiently in English, so speaking skills are quite important. Unlike traditional written English teaching, oral teaching focuses on standard pronunciation. Although the forms of teaching have diversified, spoken English teaching is still at an artificial stage at this stage. In the traditional language teaching process, teachers provide comprehensive training such as listening, reading, and writing to students through a face-to-face approach, so as to achieve the purpose of developing students' language communication skills [9–11]. Among them, the learning and training of standard spoken

language is the foundation and focus of English learning. Due to the constraints of teachers' resources, learning costs, and learning locations, the effect of traditional speaking learning and training is not satisfactory. Teachers need to spend a lot of time and effort conducting various subjective tests on students, resulting in ineffective work efficiency, especially in large-scale speaking test scenarios.

Currently, researchers are beginning to experiment with computer-assisted pronunciation training systems to address these problems [12–14]. The core issue of computer-aided pronunciation training systems is pronunciation bias testing, i.e., pronunciation bias assessment. Pronunciation bias assessment is the assessment of the standard of the learner's pronunciation and the assignment of a corresponding score or grade, which is the core function of a computer-aided pronunciation training system. Pronunciation bias assessment is mostly a confidence-based method. The phoneme sequence is first standardized and sliced to obtain more accurate phoneme boundary information. Then, the confidence of the phonemes in each speech segment is calculated, and the pronunciation bias is measured by the confidence score. Common confidence calculation methods include log-likelihood, log-likelihood ratio, log-posterior probability, and Goodness of Pronunciation (GOP) [15–17]. In addition, some methods combine confidence calculations with pronunciation features, which yield better joint score results. In order to assess pronunciation bias with high accuracy, more and more researchers are focusing on the detection of pronunciation bias at the phoneme level.

There are two ideas for the study of automatic detection of pronunciation bias at the phoneme level [18]. One is an automatic method for the detection of pronunciation bias based on acoustic phonetics. Such methods are based on a statistical analysis of speech. The other is an automatic method of pronunciation bias detection based on automatic speech recognition technology.

1.1. Pronunciation Bias Detection Based on Acoustic Phonetics. Pronunciation bias detection based on acoustic phonetics finds a specific combination of features by extracting structural, acoustic, and perceptual features of the speech to be tested. Then, pronunciation bias detection is achieved by statistically examining. A similarity calculation or a classifier is usually chosen for the differentiation of pronunciation bias types.

Morlett Paredes et al. [19] proposed a hybrid method based on time-domain features and phoneme boundary information for pronunciation bias detection of basic English pronunciation units, with remarkable results. This hybrid method used a multilayer perceptron as a classifier. Nakamura et al. [20] extracted several resonance peaks from different frames after pre-processing the speech to be measured. Then, a Gaussian Mixture Model (GMM) was used for classification and vowel articulation bias detection was achieved. Dashti and Razjmoo [21] defined a resonance peak that reduces ambient noise. This resonance peak is able to simulate the vocal tract shape properties. Articulatory bias detection is then performed by calculating the degree of structural distortion (Bhattacharyya distance) between the speech to be measured and the standard speech.

1.2. Pronunciation Bias Detection Based on Automatic Speech Recognition Technology. Automatic speech recognition is essentially a classification matching problem, while pronunciation bias detection is a classification regression problem, so pronunciation bias detection can be solved using speech recognition technology. Pronunciation bias detection based on automatic speech recognition is simpler than pronunciation bias detection based on acoustic phonetics. This is because automatic speech recognition can use a language model to counteract the effects of imprecise acoustics and thus output a legitimate sequence of characters. Therefore, this study chose to use automatic speech recognition to implement a spoken English assessment system. The key elements of automatic speech recognition technology include the extraction of speech feature parameters and the selection of acoustic models, both of which are also the focus of this study.

First of all, the extraction of speech feature parameters is a key step in the process of dynamic speech recognition, and the selection of parameters directly affects the overall performance of the system. After the speech signal has been preprocessed, it needs to be extracted and analyzed for the feature parameters. The most typical method of extraction is the use of vocoders.

The vocoder was born in the 1920s at Bell Labs in the USA. Since then, the vocoder has seen a period of rapid development. A large number of researchers have been working on speech coding and speech synthesis, and have achieved considerable results. The basis of the vocoder is Linear Predictive Coding (LPC). In the early 1980s, the US Department of Defense published LPC-10. Liu et al. [22] used LPC to build a parametric pronunciation bias database and combined it with a Gaussian Hidden Markov Model to achieve classification detection of pronunciation bias. Hiroya and Mochida [23] used LPC to extract speech feature parameters and then used the linear discriminant analysis or decision trees to train classification models to achieve pronunciation bias detection.

Second, for pronunciation detection, the constraint provided by the language model is not helpful as it leads to missed detection of incorrect pronunciations. Therefore, robust acoustic models are important to distinguish between those with standard pronunciation and those with abnormal pronunciation. In traditional speech recognition, the Gaussian Hidden Markov Model (GMM-HMM) has been the dominant acoustic model [24]. However, with the continuous development of deep learning techniques, deep learning models are gradually being used more often in speech recognition tasks. A convolutional neural network (CNN) is a multilayer perceptron that incorporates convolutional computation. CNN is one of the representative algorithms of deep learning [25] and is commonly used to analyze visual images. The CNN consists of an input layer, a convolutional layer, a ReLU activation layer, a pooling layer,

and a fully connected layer. The CNN is also known as a "translation-invariant artificial neural network."

When applied to automatic speech recognition applications, in terms of input, CNN-based automatic speech recognition techniques are broadly divided into two types: one is to use traditional acoustic feature parameters as input, such as Mel Frequency Cepstrum Coefficient (MFCC) [26], LPC [27], and Fbank [28]. The other is to use original time-frequency spectrum as input, that is, to treat the time-frequency diagram as an image. Er et al. [29] analyzed the research on deep learning techniques in speech recognition and the key problems to be solved. Nakashika et al. [30] used recurrent neural networks for speech recognition and the recognition accuracy was high.

From a pronunciation bias detection perspective, we want to retain as much of the original information as possible in the features received at the input. This is because the original information is the most realistic representation of the quality of the learner's spoken language. However, time-frequency maps can cause information loss in the frequency domain, which is detrimental to pronunciation bias detection. Therefore, the automatic speech recognition technology in this paper uses acoustic feature parameters as input information. Due to the short-time smoothness of spoken English, the feature parameters of the acoustic model in pronunciation bias detection are updated less frequently, which effectively reduces the coding bit rate (below 2.4 kb/s or even below). The simple LPC vocoder is able to achieve a range of 0.8 to 2.4 kb/s in terms of coding efficiency, which just meets the coding bit rate requirements [31–33]. Therefore, LPC is used for speech signal feature extraction, and the features are trained by convolutional neural network algorithm to complete speech recognition. The aim of this study is to adopt LPC to extract acoustic feature parameters and use CNN as an acoustic model for pronunciation bias detection to automate the detection of English pronunciation bias.

In order to achieve automatic scoring of spoken English tests, a machine learning assessment system based on linear predictive coding is proposed, which mainly consists of being divided into four modules: acoustic model acquisition module, speech recognition module, standard pronunciation transcription module, and decision module. The improved stimulated linear predictive speech coding method is used to obtain the feature parameters of the speech signal and generate the speech feature vector to implement the speech recognition module. Finally, the CNN model is used to train the speech features so as to implement the acoustic model acquisition module. The experimental results show that the improved LPC + CNN-based evaluation system can accurately detect pronunciation bias information.

The main innovations and contributions of this paper include.

(1) How accurately unvoice/voiced tones judgments are made is important for spoken English assessment systems. Therefore, the traditional LPC algorithm is improved by using hybrid excitation instead of simple binary excitation. In the acoustic feature parameter extraction process, the sub-band sound intensity of the speech signal is extracted using a split-band hybrid excitation technique in addition to the extraction of the fundamental tone period required by the traditional LPC model.

(2) An English spoken pronunciation evaluation system based on improved LPC and CNN is constructed. The improved LPC algorithm is used to obtain the feature parameters of the speech signal and generate the speech feature vector, thus realizing the speech recognition module. A CNN is used to train the speech features, thus realizing the acoustic model acquisition module.

The rest of the paper is organized as follows: In Section 2, the representative spoken pronunciation assessment system was studied in detail, while Section 3 provides the improved LPC algorithm. In Section 4, the machine learning evaluation system based on ILPC + CNN was studied in detail, while Section 5 provides experimental results and analysis. Finally, the paper is concluded in Section 6.

2. Representative Spoken Pronunciation Assessment System

Since the 1990s, many technology companies and research institutes have conducted in-depth research in the field of pronunciation bias testing and have achieved remarkable results, and launched various application systems, as shown in Table 1. These systems have been widely used in areas such as computer-aided pronunciation training, computer-aided language learning, and computer-based speaking proficiency testing. For example, the DISCO (Development and Integration of Speech technology into Courseware for language learning) project at the University of Nijmegen (Netherlands) [34]. The DISCO system automatically detects pronunciation deviations and grammatical errors in the speech to be tested and generates detailed feedback on the errors checked. The HUGO system, developed by Kyoto University in Japan for Japanese learners of English, uses a decision tree technique based on linguistics and a phonological database to check pronunciation bias.

3. Improvements to the LPC Algorithm

3.1. Principle of LPC. The most basic low-rate speech coding method is linear predictive coding. In speech signal analysis linear prediction not only enables predictive functions but also provides a very good estimation of the vocal channel model parameters. Linear prediction analysis can provide a set of speech signal model parameters that accurately represent the spectral amplitude of the speech signal. The basic idea of linear predictive analysis is to use the p sample point values of the previous set of data to predict the sample point values of the current or next set. LPC can simulate the human articulatory system very well and therefore has some advantages in the extraction of English speech feature parameters [35]. After waveform interception and noise filtering of the speech signal, multiple frames of speech signal

Table 1: Representative spoken pronunciation assessment systems.

System name	Research and development institutions
EduSpeak SDK7	Stanford research institute, USA
DISCO	Lanemegen University, The Netherlands
SCILL	University of Cambridge, UK, Massachusetts Institute of Technology, USA
TBALL	University of California, Los Angeles, USA
HUGO	Kyodo University, Japan
LISTEN	Carnegie Mellon University, USA
ISLE	University of Leeds, UK, University of Hamburg, Germany
EyeSpesk	EyeSpesk inc
Enunciate	The Chinese university of Hong Kong
PLASER	Hong Kong University of Science and Technology
National general language testing system	KDDI Corporation
Versant	Pearson Corporation

in a certain time period can be obtained by frame sampling and combined with a linear time domain model to achieve feature parameter extraction.

Let $s(n)$ represent the speech signal. According to the LPC principle, $s(n)$ can be represented by the previous p sample points.

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + +a_p s(n-p) \tag{1}$$

where a_1, a_2, a_p denote linear prediction coefficients.

Let $\hat{s}(n)$ be the predicted speech signal, then its representation is shown as follow:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

The prediction error is calculated as shown as follow:

$$\begin{aligned} e(n) &= s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \\ E &= \sum_{n} e^2(n) = \sum_{n} \left[s(n) - \sum_{k=1}^{p} a_k s(n-k) \right]^2 \end{aligned} \tag{3}$$

Let $\partial E / \partial a_k = 0 \, (1 \le k \le p)$, then all coefficients can be solved and a stable speech feature signal can be obtained.

The basic principle of the linear predictive vocoder is that the model parameters are encoded with the excitation parameters using linear predictive analysis in an all-pole vocal channel model, resulting in the transmission of high-quality speech at low bit rates (below 2.4 kb/s). The principle of the linear predictive vocoder is shown in Figure 1. At the receiver end of the linear predictive vocoder, the prediction coefficients obtained from the linear predictive analysis can be used to synthesize the transmitted speech directly [36]. Figure 2 shows the coding principle of the LPC-10 vocoder.

First, after a low-pass analog filter, the LPC-10 vocoder performs an A/D conversion at a sampling rate of 8 kHz to obtain the digitized information of the speech. The digitized speech is then processed simultaneously in two steps. (1) The excitation information is processed. After the speech has been framed, the characteristic parameters of each frame are extracted and encoded for transmission. After encoding, the fundamental tone period (Pitch) and the voiced/unvoice sign (V/UV) of each frame are obtained. The fundamental tone period is calculated using the average amplitude difference function (AMDF) method. (2) The extraction of the vocal channel parameters is processed.

Because most of the energy of the speech signal is concentrated in the low-frequency range and the power spectrum decays with frequency, the LPC needs to preprocess the speech signal first so that the power spectrum on the high frequencies can be increased, thus improving the accuracy of speech channel parameter extraction.

$$H_{pw}(z) = 1 - 0.9375 z^{-1} \tag{4}$$

where $H_{pw}(z)$ denotes the transfer function of the preprocessing filter.

3.2. Improvements to Incentive Sources. Conventional LPC algorithms use simple binary excitation sources (voiced/unvoice) to excite the synthesizer. Due to the low robustness, the quality of the speech synthesized by the binary excitation source is poor in the presence of high speech noise. Real-life English speech often has both voiced/unvoice tones, especially in noisy speech segments. Therefore, the result of the voiced/unvoice tones judgment can directly affect the quality of speech recognition. Therefore, improvement of the excitation source is important for spoken English evaluation systems. In this paper, a hybrid excitation is used instead of the traditional binary excitation, thus proposing an improved LPC algorithm (ILPC). In terms of parameter extraction, in addition to extracting the fundamental tone period required for conventional LPC, the hybrid excitation technique is also used to extract the sub-band sound intensity in the speech signal.

The steps for extracting the fundamental tone period in ILPC arithmetic are shown as follows:

Step 1: after passing the speech signal $x(n)$ through a low-pass filter at 900 Hz, the first 20 output values are removed to obtain $x'(n)$.

Step 2: find the maximum amplitude value of the first 100 samples and the maximum amplitude value of the last 100 samples of $x'(n)$ respectively. Select the smallest value as the threshold level L.

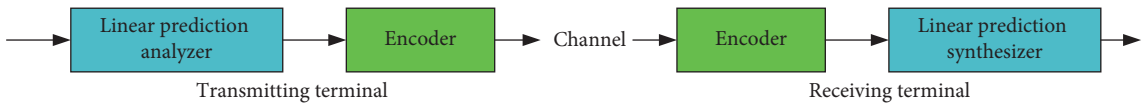

Figure 1: Linear predictive vocoder.

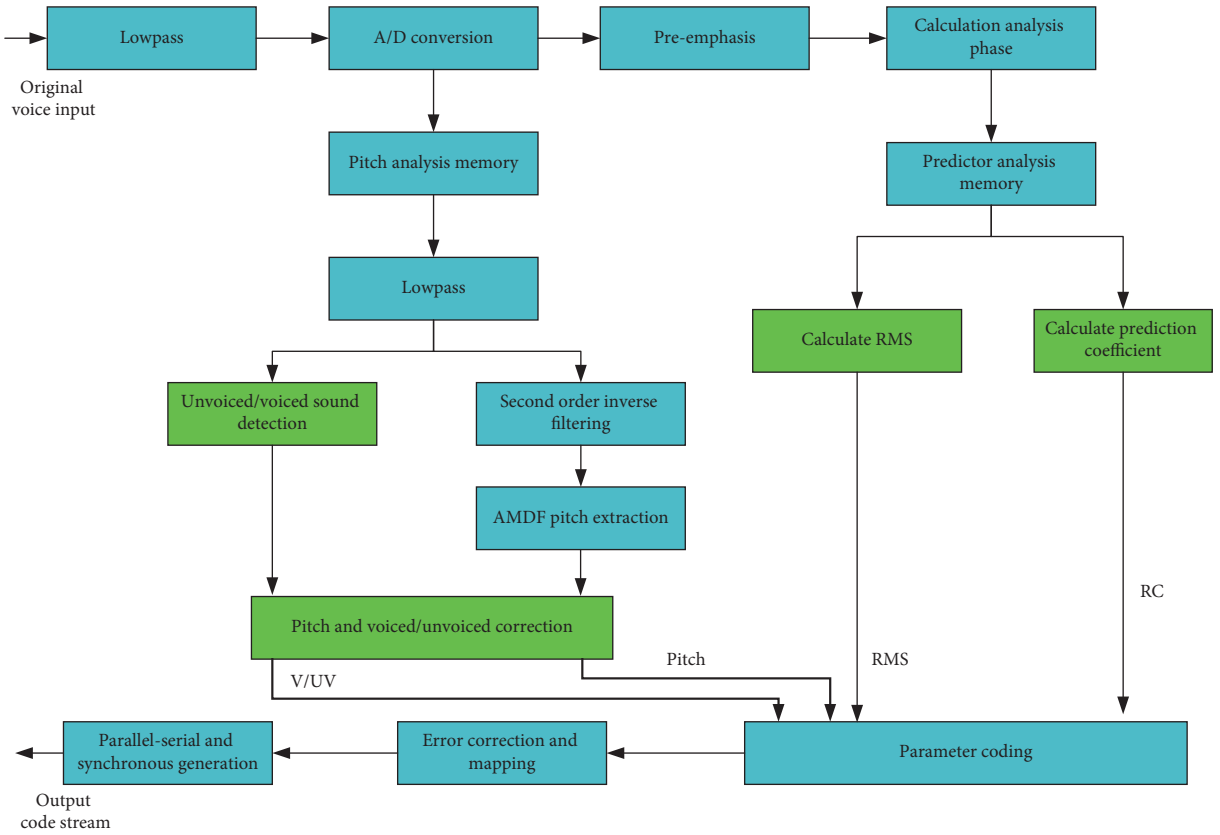

Figure 2: Encoding principle of the LPC-10 vocoder.

Step 3: make center-decimation and three-level decimation of $x'(n)$ to obtain $y(n)$ and $y'(n)$ respectively.

Step 4: find the correlation $R(k)$ between the signals $y(n)$ and $y'(n)$;

$$R(k) = \sum_{n=21}^{300} y \times y'(n+k) \tag{5}$$

where k ranges from 20 to 150 and $R(0)$ is the short-time energy.

Step 5: use the peak detector to find the maximum value of the correlation value $R_{\max}$. If $R_{\max}$ is less than $0.25R(0)$, this frame is considered as voiced tones and the fundamental tone period is set to $P = 0$. Otherwise, this frame is considered unvoiced tones, and the fundamental tone period is set to $P = k$.

The process of extracting the sub-band sound intensity in the ILPC calculation is shown in Figure 3.

After passing through the bandpass filter, the speech signal is extracted to obtain the fundamental tone period. The result of passing a frame of speech signal through each of the five sub-band filters is shown in Figure 4. The sound intensity of the five sub-bands is calculated as follows: 0.2452 for the first sub-band; 0.4478 for the second sub-band; 0.1893 for the third sub-band; 0.3707 for the fourth sub-band; and 0.3874 for the fifth sub-band.

For each unvoiced tone frame or dithered turbulent frame, the sound intensity of the speech signal in each sub-band is calculated separately. In forming the excitation signal, the sound intensity will determine the weighting of the pulse and noise sources in each sub-band, resulting in an excitation signal for the entire frequency band.

4. ILPC + CNN Based Machine Learning Evaluation System Design

4.1. General System Architecture. The automatic detection of pronunciation bias is a simulation of the human subjective detection process. Through machine learning of the manual detection results, automatic detection can even outperform human experts. The machine learning evaluation system for spoken English designed in this paper is shown in Figure 5. The system is divided into four modules: an acoustic model acquisition module, a speech recognition module, a standard pronunciation transcription module, and a decision module.

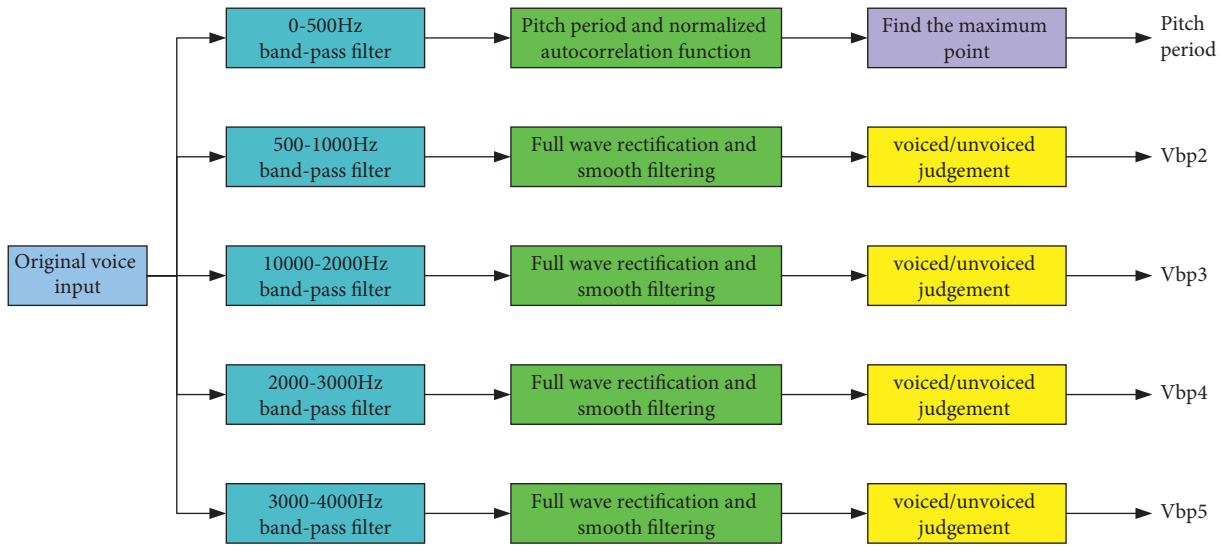

Figure 3: Extraction process of sub-band sound intensity.

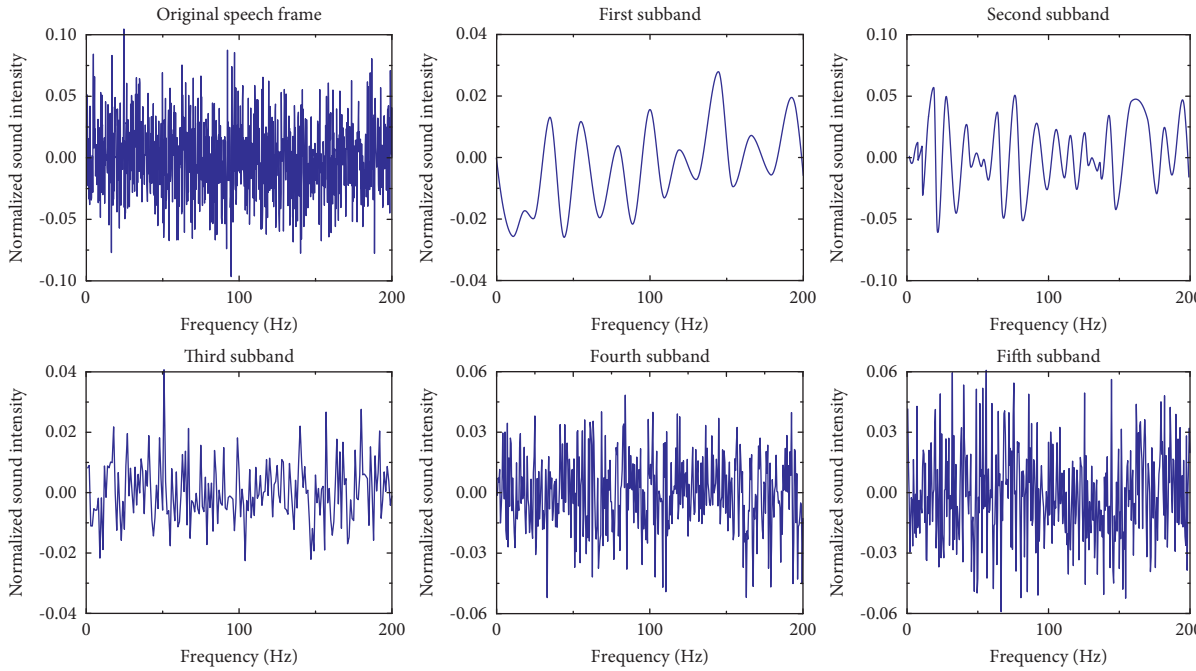

Figure 4: 5 subbands of a frame of speech.

4.2. ILPC-Based Speech Recognition Module. In this paper, the ILPC algorithm is used to implement a speech recognition module so that the learner's basic pronunciation units (phonemes), including legal and illegal pronunciation unit sequences, can be accurately identified. Automatic speech recognition aims to detect the content of the learner's pronounced text and to output legitimate character sequences by using acoustic models that can counteract the effects of undesirable acoustics. In the acoustic feature parameter extraction process, we use a split-band hybrid excitation technique to extract the sub-band sound intensity of the speech signal in addition to the parametric fundamental tone period, resulting in an accurate voiced/unvoiced tones judgment.

4.3. CNN-Based Acoustic Model Acquisition Module. The main function of the acoustic model acquisition module is to train an acoustic model. The trained acoustic model will be used in the speech recognition module. In traditional speech recognition, the Gaussian Hidden Markov Model (GMM-HMM) has been the dominant acoustic model. However, with the continuous development of deep learning techniques, deep learning models are gradually being used more often in speech recognition tasks. A convolutional neural network (CNN) is a multilayer perceptron that incorporates convolutional computation [37]. CNN is one of the representative algorithms of deep learning and is commonly used to analyze visual images. Therefore, in this paper, CNNs are used to implement the acoustic model acquisition module.

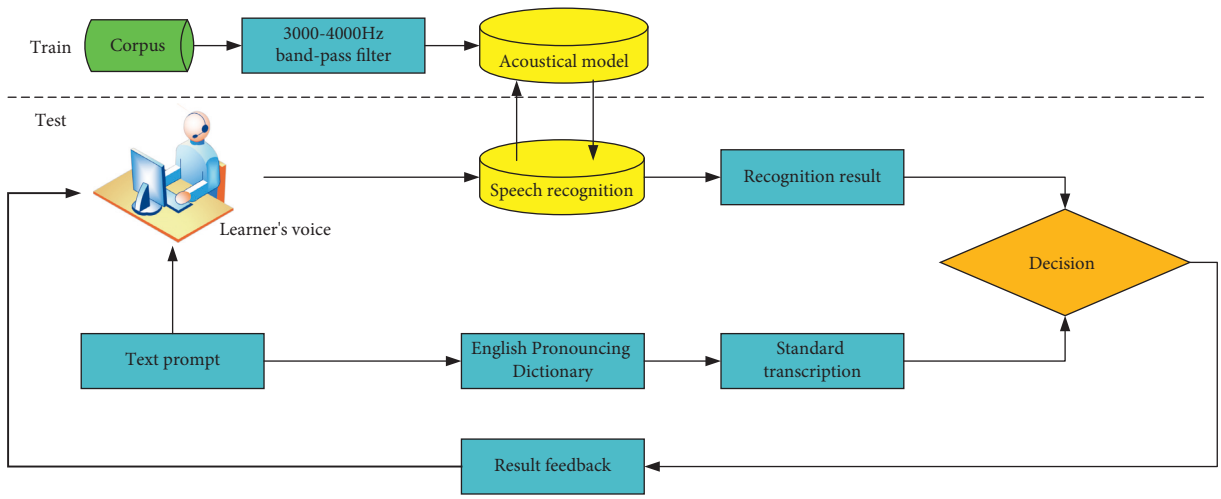

Figure 5: Overall system architecture.

The input to the CNN is the acoustic feature parameters obtained by the ILPC algorithm. the structure of the CNN is shown in Figure 6. Let the sample set of speech features be $= (x_1, x_2, x_N)$. First, the m speech features are convolved in the layer l of the CNN [38–40].

$$x_{l,j} = f\left(\sum_{j\in m} x_{l-1} * k_{lj} + b_{l,j}\right), \tag{6}$$

where k_{lj} and $b_{l,j}$ represent the weights and biases of the features j in the l layer respectively, and $*$ represents the convolution operation.

$$f(z) = \frac{1}{1+e^{-z}}. \tag{7}$$

Then, the convolution operation is performed on the m features of the N samples. Let the size of the convolution kernel be $h \times w$.

$$g(x) = \max_{1\le k\le h\times w}(x_k). \tag{8}$$

A new sample is obtained again after the convolution operation and a transformation operation is performed on it.

$$x_j^l = f\left(\sum_{i=1}^{M} a_{ij}\left(x_i^{l-1} * k_i^l\right) + b_j^l\right). \tag{9}$$

The restrictions are shown as follows:

$$\sum a_{ij} = 1, 0 \le a_{ij} \le 1. \tag{10}$$

After obtaining the fully connected layer of the convolutional neural network, the classifier is selected to predict the sample class.

In traditional acoustic model training, the label corresponding to each frame of data needs to be known in order to train effectively. Therefore, the speech signal needs to be forcibly aligned prior to training the model. Although there are some relatively mature open source alignment tools available, there are significant constraints on the performance of speech recognition techniques with forced alignment. In CNN-based acoustic models, we want to leave more tasks to the neural network to perform, such as learning how to align autonomously. Therefore, predictive alignment techniques are used to solve this problem. The loss function for predictive alignment is defined as shown below [41–43].

$$\mathrm{L}(S) = -\ln\prod(\wedge)_{(Y,\widehat{y})\in S} = -\sum_{(Y,\widehat{y})\in S}\ln p(\widehat{y}\mid Y), \tag{11}$$

where $p(\widehat{y}\mid Y)$ denotes the probability when the input sequence is $\widehat{y}$ and the output sequence is Y, and S denotes the training set. It can be seen that prediction alignment can directly output the predicted probability of a sequence without the need for external post-processing. With the help of predictive alignment, a large amount of manual resources can be saved, thus increasing the efficiency. In this paper, the acoustic model acquisition module is built by combining prediction alignment and CNN, as shown in Figure 7.

5. Experimental Results and Analysis

In order to verify the performance of ILPC + CNN in the quality assessment of spoken English, various experiments were conducted using separate speech samples with different accents. The experimental speech data were obtained from the open-source website VoxForge (https://www.voxforge.org/zh). The parameters of the experimental dataset are shown in Table 2, with a ratio of 3 : 1 between training and test samples. The parameters of the CNN model are set as shown in Table 3. The sampling rate of the audio data is 16000 Hz and the sample size was 16 bit. The number of texts (number of pronounced sentences) is 2268 and the total number of phonemes is 44359. The total number of speakers is 10 including 5 males and 5 females. First, the effect of different frame rates on ILPC performance was tested. Second, the effect of different convolutional kernel sizes on the recognition performance was tested. Finally, the designed system was compared with other spoken pronunciation evaluation systems.

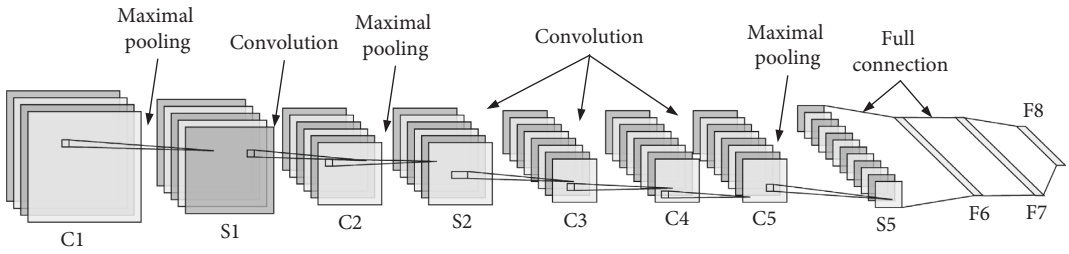

Figure 6: Structure of CNN.

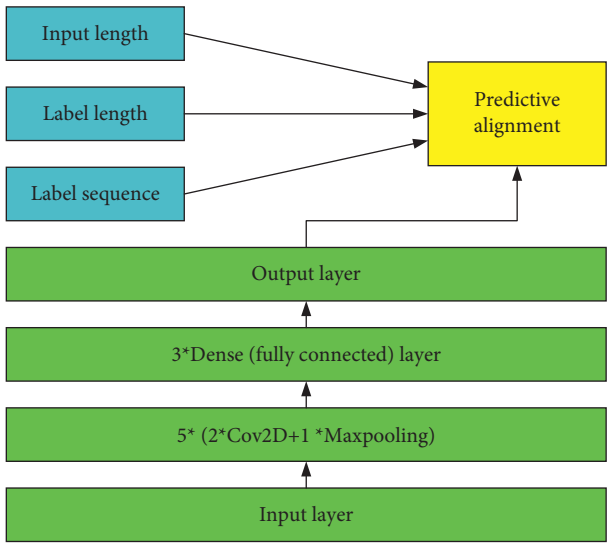

Figure 7: CNN-based acoustic model acquisition module.

Table 2: Parameters of the experimental data set.

Sample number	Type of sample	Sample size
1	American English	8276
2	British English	8144
3	European English	7768
4	Canadian English	3411
5	Australian English	2247
6	Indian English	2412

Table 3: Parameters of the CNN model.

Parameter settings	Numerical values
Learning rate	0.008
Batch_size	16
Convolution kernel	32/3 * 3
Window size	2 * 2
Droput	0.3
Epoch	300
Optimizer	Adam

5.1. Effect of Different Frame Rates on ILPC Performance. In order to obtain the best frame rate setting, the speech recognition accuracy of the six datasets at different frame rates was verified, as shown in Table 4.

It can be seen that as the number of frames extracted increases, the recognition accuracy keeps improving. As the frame rate increases to 180 Hz, the ILPC algorithm shows a high recognition accuracy. As the frame rates increases to 200 Hz, datasets 1, 2, 3, and 6 show a decrease in recognition accuracy, while datasets 4 and 5 show a very small and almost negligible increase in recognition accuracy. When ILPC was used to extract features of speech signals with different sample types, too high a frequency would increase the computational effort of speech recognition, while too low a frequency would drop important features of the speech signal. Therefore, the frame rate used in the subsequent ILPC algorithms was 180 Hz.

Table 4: Speech recognition accuracy of different frame rates.

Data sets	Frame rate (Hz)				
	60	100	150	180	200
1	0.6133	0.7491	0.8479	0.9291	0.9287
2	0.6578	0.7674	0.8667	0.9307	0.9306
3	0.6344	0.7232	0.8227	0.9267	0.9265
4	0.6473	0.7512	0.8461	0.9244	0.9246
5	0.6022	0.7334	0.8635	0.9074	0.9077
6	0.5613	0.7219	0.8218	0.9083	0.9080

5.2. Effect of Convolutional Kernel Size on Speech Recognition. To further verify the effect of convolutional kernel size on speech recognition performance, the English speech recognition accuracy under different convolutional kernel conditions was tested, as shown in Table 5.

It can be seen that the recognition accuracy of English speech decreases when the size of the convolutional kernel increases. This may be because fewer speech features are involved in the operation when the convolutional kernel size is too large, resulting in a decrease in speech recognition accuracy. The comparison shows that the recognition accuracy of CNN is higher when the convolutional kernel size is $2 * 2$ and $3 * 3$. However, $2 * 2$ is more time-consuming than $3 * 3$ in CNN operations, so to improve real-time performance, the convolutional kernel size was $3 * 3$ in subsequent experiments.

5.3. Performance Analysis of ILPC. The excitation signal of the ILPC vocoder is compared with that of the LPC vocoder, as shown in Figure 8.

It can be seen that the excitation signal obtained by ILPC is a mixed excitation signal. Each frame of speech is no longer pure unvoice tone or voice tone but contains a distinct periodic pulse string and a little noise. As a result,

Table 5: English speech recognition accuracy under different convolution kernel conditions.

Data sets	Maximum accuracy	Average accuracy rate	Standard deviation
Convolution kernel size 2 * 2			
1	0.9491	0.9291	1.75e – 003
2	0.9663	0.9307	1.92e – 003
3	0.9318	0.9267	1.66e – 003
4	0.9491	0.9244	1.73e – 003
5	0.9265	0.9074	1.81e – 003
6	0.9214	0.9083	1.65e – 003
Convolution kernel size 3 * 3			
1	0.9482	0.9286	1.77e – 003
2	0.9658	0.9301	1.96e – 003
3	0.9316	0.9259	1.73e – 003
4	0.9492	0.9238	1.72e – 003
5	0.9263	0.9066	1.82e – 003
6	0.9209	0.9078	1.66e – 003
Convolution kernel size 4 * 4			
1	0.8471	0.8312	4.13e – 003
2	0.8642	0.8476	3.37e – 003
3	0.8225	0.8164	5.68e – 003
4	0.8366	0.8132	4.05e – 003
5	0.8389	0.8246	3.49e – 003
6	0.8217	0.8059	4.22e – 003
Convolution kernel size 5 * 5			
1	0.7323	0.7191	8.24e – 003
2	0.6824	0.6162	7.93e – 003
3	0.6917	0.6694	7.27e – 003
4	0.7318	0.7161	7.13e – 003
5	0.6429	0.6225	6.83e – 003
6	0.6835	0.6634	6.65e – 003

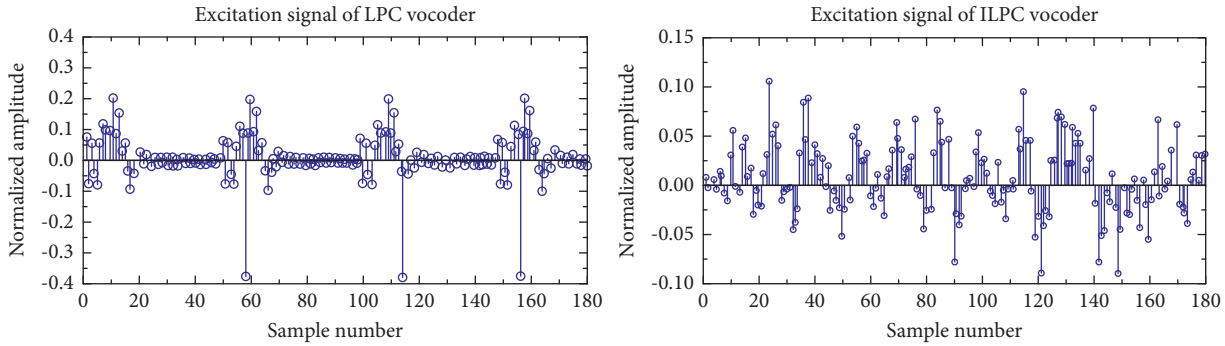

Figure 8: Comparison of excitation signals.

the speech signal obtained by ILPC is more natural and better defined. Conventional LPC uses a simple binary excitation signal to process the input sequence. Compared to the conventional LPC algorithm, ILPC based on hybrid excitation gives a waveform that more closely resembles that of the original speech signal. ILPC algorithm can get speech signals with high naturalness, and its waveform is almost consistent with the original waveform.

To verify the performance of the ILPC-based speech recognition module, 1000 samples were taken from each of the six datasets to form a speech hybrid dataset containing 6000 samples. The spoken pronunciation bias of this hybrid dataset was examined using LPC + CNN and ILPC + CNN respectively. The frame rate was 180 Hz and the convolutional kernel was 3 * 3. The detection results are shown in Table 6 and Figure 9.

It can be seen that after ILPC feature extraction, the detection accuracy of CNN is significantly improved. Due to the lower robustness, the speech quality of LPC is poor in the case of very noisy speech, which is due to the fact that real-life English speech usually has both voice tones and unvoice tones, especially in transition segments and very noisy speech segments. When using ILPC for feature extraction of the captured speech signal, each frame of speech is no longer pure voice tones and unvoice tone, thus retaining as much of the original information as possible. ILPC + CNN converges at about 140 iterations, whereas LPC + CNN takes about 180 iterations to stabilize. In

Table 6: Detection accuracy of CNN and LPC + CNN.

Algorithms	Highest recognition accuracy	Average recognition accuracy	Minimum recognition accuracy
LPC + CNN	0.8573	0.8362	0.7935
ILPC + CNN	0.9625	0.9197	0.9046

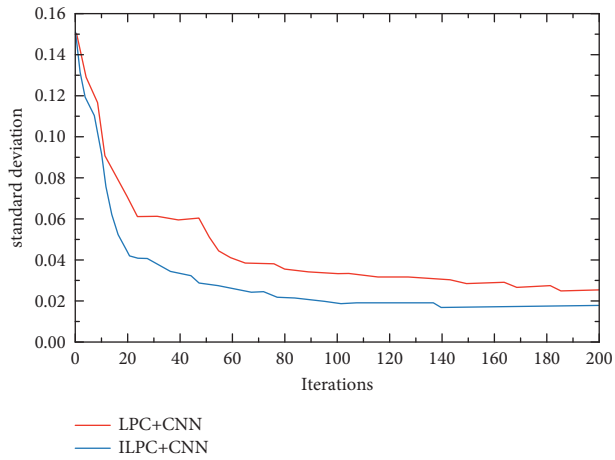

Figure 9: Standard deviation of detection for CNN and LPC + CNN.

Table 7: Articulatory phoneme detection for different systems.

System name	Number of correctly pronounced phonemes	Number of mispronounced phonemes
Manual (labelled) inspection	38326	6033
SCILL	32952	10407
TBALL	34465	9894
HUGO	32559	9568
LISTEN	34061	9714
ISLE	34954	9842
EyeSpesk	35058	8745
Enunciate	33857	9983
PLASER	36447	8243
ILPC + CNN based system	37826	7856

addition, the standard deviation of ILPC + CNN is smaller compared to LPC + CNN.

5.4. Performance Comparison of Different Spoken Language Assessment Systems. In contrast to traditional spoken pronunciation assessment methods, the training data for this experiment did not require manual annotation. Using the above speech mixture dataset containing 6000 samples, the designed system was compared with other spoken pronunciation assessment systems, the results are shown in Table 7.

A total of 44359 phonemes (initials, finals, and tones) were obtained from the speech mixture dataset. The manual detection results showed that 6033 phonemes were mispronounced in this speech data, 10407 mispronounced phonemes were detected by the SCILL system and 9894 mispronounced phonemes were detected by the TBALL system. The system designed in this paper (ILPC + CNN) detected 7856 mispronounced phonemes, which is the closest to the manual (labeled) detection result. The experimental results show that ILPC + CNN algorithm can indeed reduce the misjudgment rate of pronunciation deviation. This indicates that the feature parameters obtained by ILPC using hybrid excitation reflect well the characteristics of the original speech signal and therefore the decoded speech quality is better and the speech is clearer.

Finally, the experiment classified the 64 pronunciation errors into three types, namely initial errors, final errors, and tone errors. These three types of pronunciation errors were counted and the results are shown in Figure 10.

As you can see, of the 3 types of pronunciation errors, intonation is the most likely to occur. Therefore, learners of English need to focus on intonation. The next problem is rhyme errors. Compared to the other two types of pronunciation errors, vowel errors are easier to solve. The phenomenon that tone errors are much higher than the other two types of errors is in line with linguistic laws and therefore the experimental results are reliable.

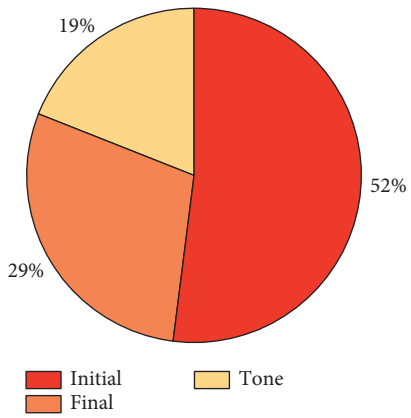

Figure 10: Percentage of 3 types of pronunciation errors.

6. Conclusion

In this paper, a machine learning evaluation system for spoken English based on ILPC + CNN algorithm is constructed so as to automate the detection of learners' pronunciation errors. The designed system consists of four main modules: acoustic model acquisition module, speech recognition module, standard pronunciation transcription module, and decision module. The speech recognition module uses the ILPC algorithm to obtain the feature parameters of the speech signal and generate the speech feature vector. The acoustic model acquisition module uses a CNN model to train the speech features and the input to the CNN is the acoustic feature parameters obtained by the ILPC algorithm. The experimental results show that the feature parameters obtained by ILPC using hybrid excitation reflect the characteristics of the original speech signal very well, and therefore the decoded speech quality is better and the speech is clearer. Compared with other spoken English evaluation systems, the ILPC + CNN-based machine learning evaluation system can reduce the misjudgment rate of pronunciation bias.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest to report regarding the present study.

References

[1] W. Y. Kim, G. S. Lee, Y. H. Kim et al., "Immunolocalization of pki α, βI, βII and γ in adult and developing rat kidney," *Electrolyte & Blood Pressure: E & BP*, vol. 5, no. 2, pp. 75–88, 2007.

[2] Z. Gang, "Quality evaluation of English pronunciation based on artificial emotion recognition and Gaussian mixture model," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 4, pp. 7085–7095, 2021.

[3] J. Kim and D. A. Craig, "Validation of a videoconferenced speaking test," *Computer Assisted Language Learning*, vol. 25, no. 3, pp. 257–275, 2012.

[4] E. D. Quaid, "Reviewing the IELTS speaking test in East Asia: theoretical and practice-based insights," *Language Testing in Asia*, vol. 8, no. 1, pp. 2–9, 2018.

[5] H. T. D. Huang and S. T. A. Hung, "Comparing the effects of test anxiety on independent and integrated speaking test performance," *Tesol Quarterly*, vol. 47, no. 2, pp. 244–269, 2013.

[6] S. Karim and N. Haq, "An assessment of IELTS speaking test," *International Journal of Evaluation and Research in Education*, vol. 3, no. 3, pp. 152–157, 2014.

[7] B. Bridgeman, D. Powers, E. Stone, and P. Mollaun, "TOEFL iBT speaking test scores as indicators of oral communicative language proficiency," *Language Testing*, vol. 29, no. 1, pp. 91–108, 2012.

[8] T. P. Stones, "Transcription and the IELTS speaking test: facilitating development," *ELT Journal*, vol. 67, no. 1, pp. 20–30, 2013.

[9] H. T. D. Huang, S. T. A. Hung, and L. Plakans, "Topical knowledge in L2 speaking assessment: c," *Language Testing*, vol. 35, no. 1, pp. 27–49, 2018.

[10] F. Nakatsuhara, C. Inoue, and L. Taylor, "Comparing rating modes: analysing live, audio, and video ratings of IELTS speaking test performances," *Language Assessment Quarterly*, vol. 18, no. 2, pp. 83–106, 2021.

[11] P. Seedhouse, "The dual personality of 'topic' in the IELTS Speaking Test," *ELT Journal*, vol. 73, no. 3, pp. 247–256, 2019.

[12] J. Li, "An evaluation of IELTS speaking test," *OALib*, vol. 06, no. 12, pp. 1–17, 2019.

[13] K. Evanini, M. Heilman, X. Wang, and D. Blanchard, "Automated scoring for the*TOEFL Junior*® Comprehensive writing and speaking test," *ETS Research Report Series*, vol. 2015, no. 1, pp. 1–11, 2015.

[14] S. Roshan, "A critical review of the revised IELTS speaking test," *International Journal of English Language Education*, vol. 2, no. 1, p. 120, 2013.

[15] A. Latifa, A. Rahman, A. Hamra, B. Jabu, and R. Nur, "Developing a practical rating rubric of speaking test for university students of English in parepare, Indonesia," *English Language Teaching*, vol. 8, no. 6, pp. 166–177, 2015.

[16] Y. Zhan and Z. H. Wan, "Test takers' beliefs and experiences of a high-stakes computer-based English listening and speaking test," *RELC Journal*, vol. 47, no. 3, pp. 363–376, 2016.

[17] S. O'Grady, "The impact of pre-task planning on speaking test performance for English-medium university admission," *Language Testing*, vol. 36, no. 4, pp. 505–526, 2019.

[18] R. Koizumi and A. Hirai, "Comparing the story retelling speaking test with other speaking tests," *JALT Journal*, vol. 34, no. 1, p. 35, 2012.

[19] A. Morlett Paredes, A. Gooding, L. Artiola I Fortuny et al., "The state of neuropsychological test norms for Spanish-speaking adults in the United States," *The Clinical Neuropsychologist*, vol. 35, no. 2, pp. 236–252, 2021.

[20] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[21] L. Dashti and S. A. Razmjoo, "An examination of IELTS candidates' performances at different band scores of the speaking test: a quantitative and qualitative analysis," *Cogent Education*, vol. 7, no. 1, Article ID 1770936, 2020.

[22] P. Liu, S. Li, and H. Wang, "Steganography integrated into linear predictive coding for low bit-rate speech codec," *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 2837–2859, 2017.

[23] S. Hiroya and T. Mochida, "Speech sound naturalness alters compensation in response to transformed auditory feedback," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, p. 3228, 2016.

[24] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," *Speech Communication*, vol. 49, no. 6, pp. 501–513, 2007.

[25] F. Demir, M. Turkoglu, M. Aslan, and A. Sengur, "A new pyramidal concatenated CNN approach for environmental sound classification," *Applied Acoustics*, vol. 170, no. 6, Article ID 107520, 2020.

[26] Y. C. Zeng, Y. Y. Chen, and Y. H. Mao, "Mel frequency Cepstrum coefficient extraction method based on empirical mode decomposition and combined spectrum of fourier transform and wigner distribution," *Natural Science Journal of Xiangtan University*, vol. 132, no. 5, pp. 563–573, 2015.

[27] Y. Hiwasaki, K. Mano, and T. Kaneko, "An LPC vocoder based on phase-equalized pitch waveform," *Speech Communication*, vol. 40, no. 3, pp. 277–290, 2003.

[28] J. Cao, M. Cao, J. Wang, C. Yin, D. Wang, and P. P. Vidal, "Urban noise recognition with convolutional neural network," *Multimedia Tools and Applications*, vol. 78, no. 20, pp. 29021–29041, 2019.

[29] M. B. Er, E. Isik, and I. Isik, "Parkinson's detection based on combined CNN and LSTM using enhanced speech signals with Variational mode decomposition," *Biomedical Signal Processing and Control*, vol. 70, Article ID 103006, 2021.

[30] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 580–587, 2015.

[31] A. Harma and U. K. Laine, "A comparison of warped and conventional linear predictive coding," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 579–588, 2001.

[32] R. J. Javier and Y. Kim, "Application of linear predictive coding for human activity classification based on micro-Doppler signatures," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 10, pp. 1831–1834, 2014.

[33] M. W. Spratling, "A review of predictive coding algorithms," *Brain and Cognition*, vol. 112, pp. 92–97, 2017.

[34] D. Bosbach and L. Z. Evins, "The European DISCO project: deep geological disposal of modern spent nuclear fuel," *Safety of Nuclear Waste Disposal*, vol. 1, pp. 233-234, 2021.

[35] M. F. Anjum, S. Dasgupta, R. Mudumbai, A. Singh, J. F. Cavanagh, and N. S. Narayanan, "Linear predictive coding distinguishes spectral EEG features of Parkinson's disease," *Parkinsonism & Related Disorders*, vol. 79, pp. 79–85, 2020.

[36] S. K. Selvaperumal, C. Nataraj, and V. Thiruchelvam, "Speech to text synthesis from video automated subtitling using Levinson Durbin method of linear predictive coding," *International Journal of Applied Engineering Research*, vol. 11, no. 4, pp. 2388–2395, 2016.

[37] A. Pandey and D. L. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[38] S. Kwon and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, pp. 183–201, 2019.

[39] V. Passricha and R. K. Aggarwal, "A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition," *Journal of Intelligent Systems*, vol. 29, no. 1, pp. 1261–1274, 2019.

[40] Z. Zhao, Q. Li, Z. Zhang, N. Cummins, H. Wang, and J.. Tao, "Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition," *Neural Networks*, vol. 141, pp. 52–60, 2021.

[41] J. Zhao, X. Mao, and L. Chen, "Learning deep features to recognise speech emotion using merged deep CNN," *IET Signal Processing*, vol. 12, no. 6, pp. 713–721, 2018.

[42] L. Comanducci, P. Bestagini, M. Tagliasacchi, A. Sarti, and S Tubaro, "Reconstructing speech from CNN e," *IEEE Signal Processing Letters*, vol. 28, pp. 952–956, 2021.

[43] X. Chen, "Simulation of English speech emotion recognition based on transfer learning and CNN neural network," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 2349–2360, 2021.