IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 5, October, 2024, pp. 2751-2761
Journal homepage: http://iieta.org/journals/ts

An English Pronunciation Correction System Based on Signal Processing

Jin Gan

College of General Education, Chengdu Jincheng College, Chengdu 611731, China

Corresponding Author Email: ganjin@cdjcc.edu.cn

https://doi.org/10.18280/ts.410548

Received: 8 March 2024
Revised: 30 July 2024
Accepted: 15 September 2024
Available online: 31 October 2024

Keywords:
English pronunciation correction, signal processing, English speech recognition, personalized feedback, language learning system

ABSTRACT

With the advancement of globalization, the importance of English as a global lingua franca has grown, making standard English pronunciation a reflection of personal communication skills and international competitiveness. However, non-native English speakers often struggle with pronunciation due to the influence of their mother tongue, which negatively impacts communication effectiveness. To address this, an English pronunciation correction system based on signal processing technology has emerged, aimed at helping learners improve pronunciation accuracy and enhance language learning efficiency. Nonetheless, existing systems still face challenges in areas such as processing complex English speech signals, recognition accuracy, real-time performance, and adaptability. This research proposes two main contributions: the construction of an English speech recognition network based on signal processing to improve recognition accuracy and real-time performance, and the development of a pronunciation correction method that offers precise and personalized feedback. This study not only addresses the shortcomings of current systems but also provides new ideas and methods for the future development of pronunciation correction systems, holding significant theoretical and practical value.

1. INTRODUCTION

With the acceleration of globalization, the importance of English as a global lingua franca has become increasingly prominent [1-5]. Mastering standard English pronunciation is not only an important reflection of personal communication skills but also an expression of national and corporate competitiveness on the international stage. However, many non-native English speakers often experience interference from their mother tongue's phonetic system when learning English, leading to inaccurate pronunciation that affects communication effectiveness [6, 7]. Addressing this issue, how to utilize modern technological means, especially signal processing technology, to assist in the correction of English pronunciation has become an important topic in language learning research.

In the past few decades, researchers have conducted extensive studies and explorations on English pronunciation correction systems. By analyzing and correcting English pronunciation using signal processing technology, it is possible not only to improve learners' pronunciation accuracy but also to enhance their language learning efficiency to some extent [8-10]. Therefore, English pronunciation correction systems based on signal processing hold significant research significance in the field of educational technology, providing language learners with more scientific and effective pronunciation correction tools.

Although some signal processing-based pronunciation correction systems have already been developed, these systems still exhibit certain deficiencies in practical applications [11-16]. For example, most existing systems have low recognition accuracy and real-time performance when processing complex English speech signals, making it difficult to provide effective personalized pronunciation correction feedback. Additionally, these systems also need to improve their adaptability and robustness when handling English speech signals with different accents and speech rates [17-23]. Thus, addressing these deficiencies and further enhancing system performance and learner experience remains an important issue that requires urgent resolution.

This paper focuses on two main research components: first, constructing an English speech recognition network based on signal processing, aimed at improving the recognition accuracy and real-time performance of English speech signals; second, researching signal processing-based English pronunciation correction methods to provide learners with more precise and personalized pronunciation correction suggestions. This study not only fills the gaps in recognition accuracy and real-time performance of existing systems but also provides new ideas and methods for the future development of English pronunciation correction systems, possessing significant theoretical and practical value.

2. CONSTRUCTION OF THE ENGLISH SPEECH RECOGNITION NETWORK BASED ON SIGNAL PROCESSING

An efficient English speech recognition network is crucial for accurately capturing and decoding learners' English speech

signals, which is essential for pronunciation analysis and correction. Without precise English speech recognition, the system cannot correctly understand learners' pronunciation features, thereby failing to provide effective correction suggestions. English speech contains rich acoustic features and complex pronunciation variations, posing significant challenges for English speech recognition.

To address the problem of long-distance dependencies, traditional English speech sequence-to-sequence models often encounter information forgetting when processing long sequences, which creates significant obstacles in accurately capturing learners' English speech features. Models based on multi-head attention mechanisms better handle global information by assigning different weights to various elements, allowing the model to focus on parts of the input sequence that contribute more to the output, thereby alleviating the long-distance dependency issue to some extent.

To better tackle long-distance dependencies and improve the accuracy and robustness of English speech recognition, this paper employs a network model based on dilated convolution and multi-head attention mechanisms for recognizing English speech, providing a solid technical foundation for the English pronunciation correction system. The structure diagram is shown in Figure 1.

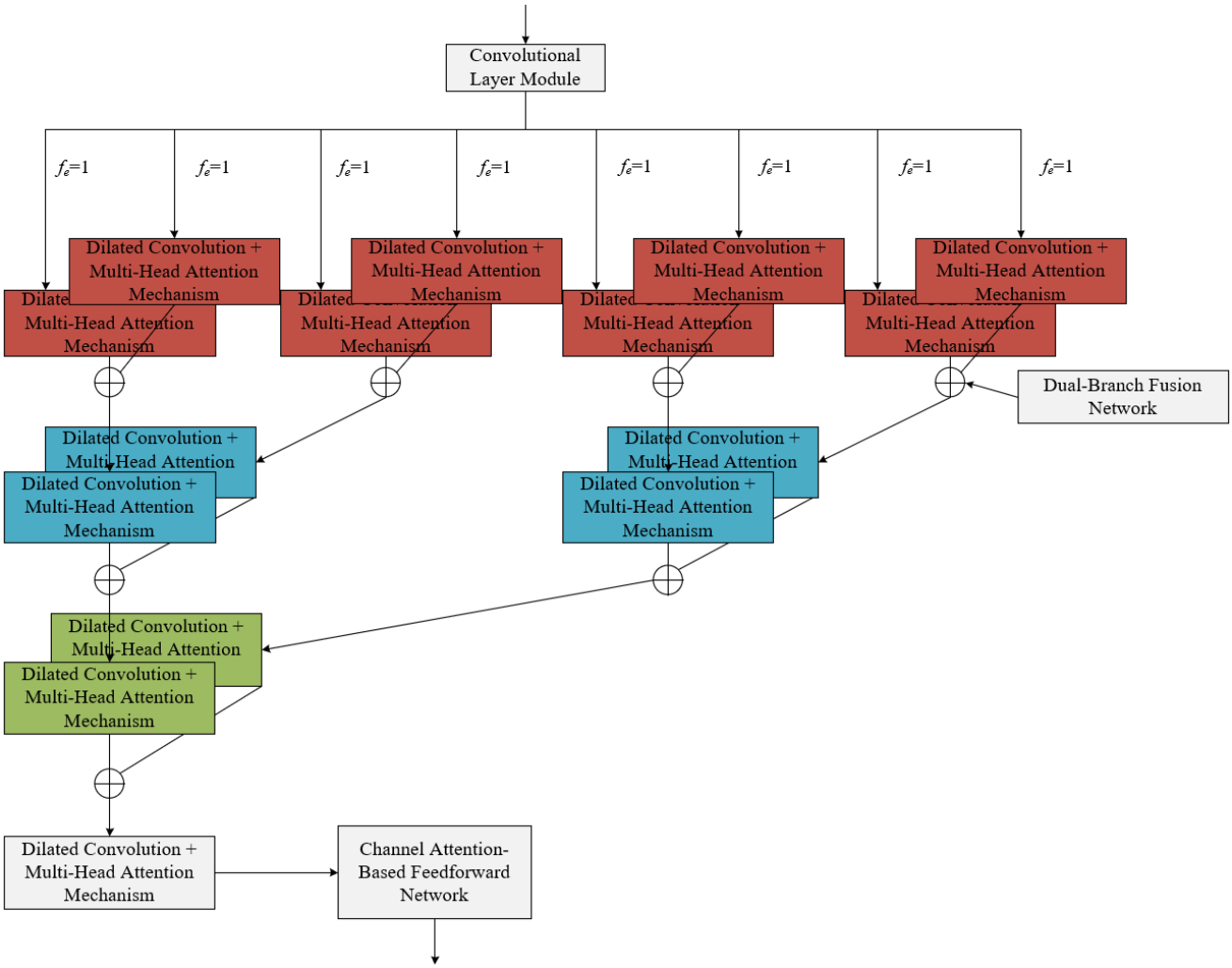

Figure 1. Structure of the English speech recognition model

To effectively extract shallow local features from English speech signals and ensure the generation of denser and more representative English speech feature vectors, we add a convolutional layer module before the network model. This module mainly consists of a convolutional downsampling module, a positional encoding module, and a one-dimensional convolution module. The convolutional downsampling module extracts local features through two two-dimensional convolution networks. The input English speech signal is a four-dimensional vector with an added channel, denoted as (Y, S, D, Z), where Y represents batch size, S represents the number of frames in a sample, D represents the feature dimension of each frame signal, and Z represents the number of added channels. The convolutional downsampling module reduces the dimensionality of the input English speech signal and merges its last two dimensions, resulting in a dimensional change to (Y, S, DZ). This processing transforms the original four-dimensional vector into a two-dimensional feature sequence (Y, S, DZ), thereby enabling more efficient subsequent feature extraction.

After the convolutional downsampling, adding positional encoding helps the network better capture the positional relationships within the sequence information. This is especially critical for the time series features in English speech signals, as they contain not only frequency information but also dynamic changes over time. By incorporating positional encoding, the network can more accurately understand the temporal distribution of the English speech signal, thus enhancing the feature extraction effect. In practical implementation, the positional encoding module generates a

positional encoding vector with the same dimensions as the input English speech feature sequence. Assuming the input sequence is $a=[a_1,...,a_V]$ with length V, where v denotes the position of elements in the sequence, a common method for generating positional encoding is based on sine and cosine functions, which can capture the relative positional relationships of elements in the sequence. The formula for generating the encoding vector at position v is as follows:

$$\vec{o}_v^{(u)} = d(v)^{(u)} = \begin{cases} SIN(\mu_j \cdot v) u = 2j, \\ COS(\mu_j \cdot v) u = 2j+1. \end{cases} \tag{1}$$

Assuming that the positional encoding vector at even positions is represented by $u=2j$ and at odd positions by $u=2j+1$, the frequency μ_j is expressed as follows:

$$\mu_j = \frac{1}{1000^{2j/f}} \tag{2}$$

The vector form of the positional encoding is given by:

$$\vec{o}_n = \begin{bmatrix} SIN(\mu_1 \cdot v) \\ COS(\mu_1 \cdot v) \\ SIN(\mu_2 \cdot v) \\ COS(\mu_2 \cdot v) \\ \vdots \\ SIN(\mu_f / 2 \cdot v) \\ COS(\mu_f / 2 \cdot v) \end{bmatrix} \tag{3}$$

By adding the above vector to the English speech feature vector through addition, we have:

$$\tilde{a}_n = a_n + \vec{o}_n \tag{4}$$

The one-dimensional convolution module further processes the two-dimensional feature sequence. Through one-dimensional convolution operations, the network can extract features along the temporal dimension, capturing finer-grained English speech features. Its structure includes layer normalization, pointwise convolution, channel convolution, gated linear units (GLU), batch normalization, Swish activation function, Dropout, and residual structures. Layer normalization and batch normalization enhance training stability and accelerate convergence within the one-dimensional convolution module. The combination of pointwise convolution and channel convolution allows the one-dimensional convolution module to efficiently extract and process English speech features. GLU introduces a sequential processing capability similar to Recurrent Neural Networks in the one-dimensional convolution module. The Swish activation function and Dropout further improve the model's nonlinear expressiveness and prevent overfitting. The application of residual structures in the one-dimensional convolution module ensures the training stability and performance enhancement of deep networks.

The basic principle of the convolutional layer module is to efficiently process and represent the input English speech signals by learning and extracting different local features from English speech signals through local connections and weight sharing using multiple convolution kernels. When the input signal $a(v)$ enters the convolutional layer module, it first undergoes layer normalization to stabilize the distribution of data, reducing internal covariate shift during training. The normalized signal is convolved with a convolution kernel $q(v)$ of size j and quantity $re*R$, where R is the embedding dimension of the attention mechanism. The convolution features A obtained through the convolution operation effectively extract local features from the input signal. The calculation formula is as follows:

$$A = a(v) * q(v) = \sum_{l=-j}^{j} a(l) \cdot q(v-l) \tag{5}$$

Afterwards, the convolution features A are fed into the GLU. The GLU compresses and processes all information prior to the current moment within the time window, maintaining the consistency between the temporal position and the actual content. In this way, the GLU can retain the temporal information in the English speech signal while processing information from different time points in parallel, which is crucial for handling the temporal features in English speech signals. Assuming the convolution kernel parameters are represented by Q and N, and the bias parameters are represented by y and z, the calculation formula is as follows:

$$g(A) = (A*Q+y) \otimes \delta(A*N+z) \tag{6}$$

The feature map undergoes batch normalization, ensuring that the features conform to a standard normal distribution, which further improves the model's training stability and convergence speed. Batch normalization reduces internal covariate shift, ensuring consistency in the distribution of input data across layers, thus enhancing the network's generalization capability. When the normalized features are input into the Swish activation layer, let G denote the features after activation by the Swish function, with batch normalization denoted as $BN(.)$ and the Swish activation function denoted as $\delta(.)$. The calculation formula is:

$$G = \delta\left(BN\left(fq\ CONV(g(A))\right)\right) \tag{7}$$

The expression for $\delta(.)$ is:

$$\delta(a) = \frac{a}{1+e^{-a}} \tag{8}$$

The English speech features G then pass through a pointwise convolution and a Dropout layer. The pointwise convolution is used for cross-channel feature fusion, combining features from each channel using a 1×1 convolution kernel, further extracting and merging feature information across different channels. Dropout randomly drops some neurons to prevent overfitting and improve the model's generalization capability. Let the output of the pointwise convolution be denoted by $oq\ CONV(.)$. The formula is as follows:

$$C = Dropout(oq\ CONV(G)) \tag{9}$$

The processed feature map passes through a residual structure, where the input is directly added to the output via a skip connection. This not only preserves the original feature information but also enhances the training stability of the deep network, avoiding the gradient vanishing problem, thereby improving model performance. Let the output of the stacked module be denoted by B and the weighting factor of the residual structure by β. The calculation formula is:

$$B = C + \beta \cdot a(v) \tag{10}$$

As English speech signals are temporal data, their features depend not only on the current input but are also closely related to past historical features. Therefore, the network must capture both local and global features during feature extraction. Dilated convolution expands the receptive field while maintaining computational efficiency, allowing it to capture long-range dependencies, making it suitable for extracting global features. Meanwhile, the convolutional layer effectively extracts local features from English speech signals through local connections and weight sharing. Therefore, using a combination of dilated convolution and standard convolution can effectively balance local and global information during feature extraction.

To further enhance feature extraction effectiveness, a multi-head attention mechanism is introduced into the network. This mechanism allocates multiple attention subspaces to achieve multi-scale feature learning. Specifically, the attention space is divided into V attention subspaces, with each subspace's model dimension reduced to R/V, where R represents the embedding dimension of the attention mechanism. This partitioning method maintains the total parameter count of the model, but it can somewhat diminish the feature representation capacity within each attention subspace. Therefore, it is crucial to set the number of attention heads V appropriately. An excessive number of attention heads can lead to excessively small dimensions in each subspace, resulting in insufficient representation capacity, thus affecting overall performance. When designing the English speech recognition network, using a combined model of dilated convolution and multi-head attention effectively addresses these issues.

Specifically, the dilated convolutional neural network undertakes the initial feature extraction task in each parallel branch. Since dilated convolution can significantly expand the receptive field without increasing computational cost, it can capture a larger contextual information range with fewer layers. Each branch has a different dilation rate, meaning that each branch can extract information from different scales of the feature space. This multi-scale feature extraction approach ensures that the network can focus on local details while also capturing long-range dependencies, providing good coverage of both global and detailed features in English speech signals. After the dilated convolution processing, the feature vectors from each branch are input into the multi-head attention mechanism. The multi-head attention mechanism divides the attention space into multiple subspaces, where each subspace independently computes attention weights, thus allowing focus on different features. Due to the different dilation rates in each branch, the receptive fields that the multi-head attention mechanism focuses on also vary, allowing each attention head to concentrate on information within its receptive field without overly focusing on global information. This design not only reduces the amount of information that needs to be processed in each branch but also enhances the overall feature extraction efficiency and accuracy of the network.

To address the issue of increased model parameters resulting from the multi-branch structure, this paper proposes a dual-branch fusion network, with a structure diagram shown in Figure 2. In this module, the outputs of the two parallel branches are concatenated along the channel dimension to form a new feature map. The concatenated feature map's channel dimension increases to twice the original, but a linear layer compresses it back to the original channel count, ensuring that the model parameters do not significantly increase. Layer normalization is performed after each concatenation to ensure stability in gradients during backpropagation, further improving training effectiveness and model convergence speed. Specifically, let the output of the u-th layer and the k-th branch's module be denoted as P^k_s, where u=1,2,…,v and k=1,2,…,l/2. Each layer's branches gradually merge after passing through the dual-branch fusion network until a single branch is ultimately formed. Let v denote the maximum number of layers in the network, l denotes the maximum number of branches, and the concatenation output of the u+1-th layer and the k-th branch be denoted as $CON^{u+1}{}_k$, and the network output of the u+1-th layer and the k-th branch be denoted as $FDV^k{}_{u+1}$. The output of the dual-branch fusion network is given by the following equation:

$$CON^k_{u+1} = CONCAT\left(P_u^{2\times k-1}, P_u^{2\times k}\right) \tag{11}$$

$$FDV^k_{u+1} = Dropout\left(BN\left(LINEAR\left(LN\left(CON^k_{u+1}\right)\right)\right)\right) \tag{12}$$

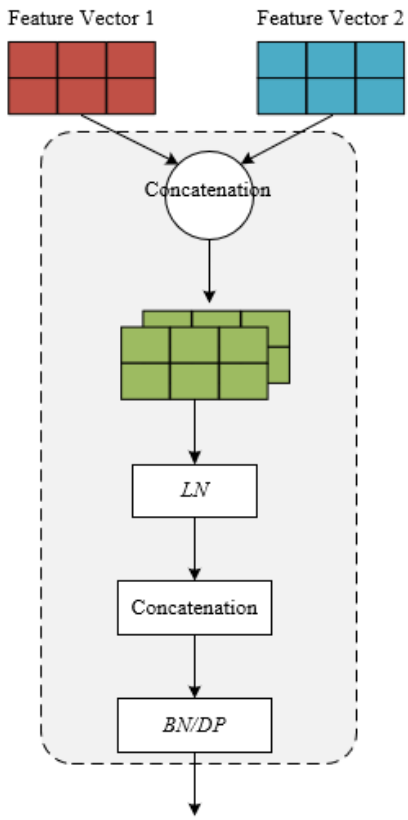

Figure 2. Structure of dual-branch fusion network

The channel attention mechanism enhances the model's ability to identify different channel features by weighting the channel dimension of the output feature map. Specifically, this

paper employs SENet to achieve this purpose. SENet computes the importance weights of each channel by performing "squeeze" and "excitation" operations on each channel feature, and applies weighted processing to the output features. This mechanism allows the model to focus more on the features that significantly contribute to the pronunciation correction task while suppressing noise or irrelevant features. The specific structure is shown in Figure 3.

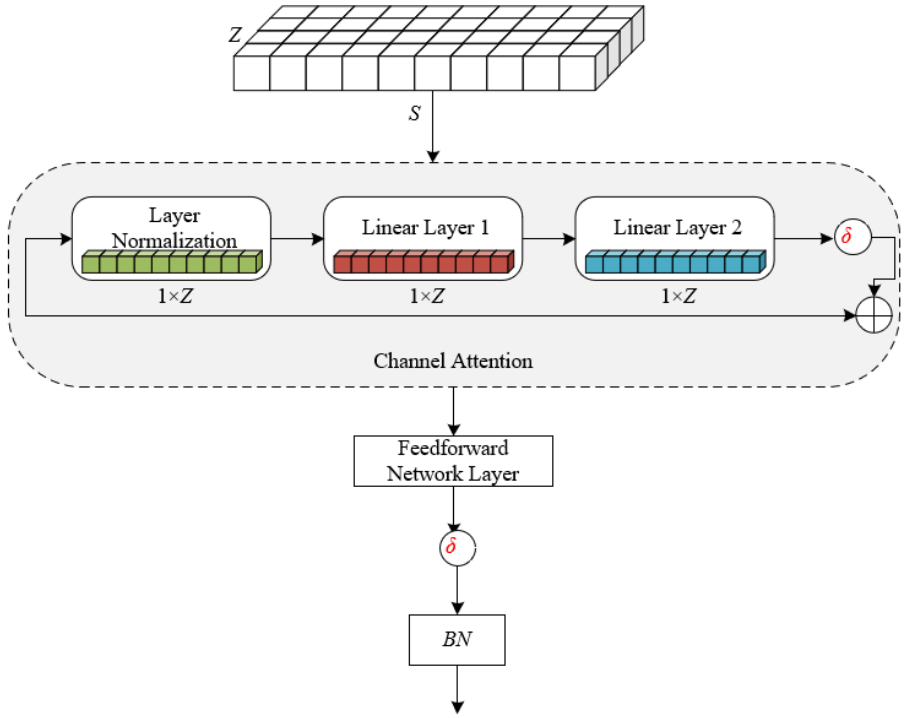

Figure 3. Structure of feedforward neural network based on channel attention

In the network model, the output dimension of the linear layer in the last attention mechanism module is set to $F=2*Z$. This design enhances the representation capability of SENet by increasing the number of channel dimensions. Through this method, the network can more flexibly capture complex English speech features, especially excelling in handling long-distance dependencies and multi-scale features. After the channel features are weighted, the feature map is input into the feedforward neural network module. The design of this module aims to further enhance the model's nonlinear representation ability and classification performance. The structure of the feedforward neural network module includes layer normalization, two linear layers, Dropout, and residual connections. Specifically, layer normalization is used to standardize the input of each layer, avoiding gradient vanishing or explosion issues, ensuring training stability. The linear layers map the input space to the output space through the stacking of two linear layers. This design not only increases the model's nonlinear representation ability but also better captures complex English speech features. Dropout is used to prevent overfitting by randomly dropping some neurons, enhancing the model's generalization ability. Residual connections ensure smoother information transfer within the network, avoiding information loss and improving the model's training efficiency.

3. RESEARCH ON SIGNAL PROCESSING-BASED ENGLISH PRONUNCIATION CORRECTION METHOD

Based on the generated signal processing English speech recognition results, preprocessing and feature extraction are performed, and these features are input into the feedforward neural network based on channel attention. A confusion matrix for pronunciation errors is constructed, followed by the application of factor clustering algorithms to analyze the confusion matrix. Based on the results of the clustering analysis, we build an automated pronunciation error detection and correction system. This system can provide targeted correction suggestions and guidance based on different types of pronunciation errors.

In the proposed signal processing-based English pronunciation correction method, we combine the Hidden Markov Model (HMM) to establish and train the learners' spoken pronunciation model. HMM is widely used in English speech recognition, effectively handling time series data and modeling the temporal features in English speech signals. Specifically, we first establish an initial model ε and initialize the parameters (g, m, τ). Here, g represents the state transition matrix, describing the transition probabilities between different states; m represents the observation matrix, describing the probability distribution of observing specific features in each state; and τ represents the initial state distribution, describing the probability distribution of each state at the initial time in the HMM. During initialization, we can use a large amount of correctly pronounced English speech data to estimate the initial values of these parameters through statistical methods.

After establishing the initial model, we obtain new parameter combinations (g^*,m^*,τ^*) through observed sequence data. In this step, we use the learners' actual pronunciation data to iteratively update the model parameters through the HMM training algorithm. Specifically, the Baum-Welch algorithm is used within the expectation-maximization framework to iteratively optimize the model parameters,

maximizing the likelihood of the model for the observed data. In each iteration, the expectation step calculates the expected value of the observed data under the current parameters, while the maximization step updates the model parameters based on this expected value.

The above steps are repeated to improve the model parameters until the iteration ends. This process typically requires multiple iterations until the changes in model parameters stabilize, i.e., the parameters converge. During this process, the model is continuously optimized, gradually improving its fit to the learners' pronunciation data. To avoid overfitting, cross-validation methods are commonly employed to evaluate the model. After each iteration, performance metrics on the validation set are used to monitor the training effectiveness of the model and prevent it from overfitting to the training data.

For each frame of the English speech signal, we use Fast Fourier Transform (FFT) to convert the time-domain signal into the frequency domain. In the frequency domain, the spectral range of the signal usually contains rich English speech information. To further analyze these spectral features, we set up several band-pass filters. These filters decompose the spectral signal to extract features within different frequency ranges. Specifically, we need to set several band-pass filters within the spectral range, evenly distributing their center frequencies. After setting up the band-pass filters, the output signal from each filter is processed. First, the spectral signal of each frame is passed through all the established band-pass filters. Each filter retains only the signal components near its center frequency, while other frequency components are attenuated. Thus, we obtain a series of signals, each representing the energy distribution within different frequency ranges. For each band-pass filter's output signal, we calculate its energy. This can be achieved by summing the squares of the filtered signal. The energy value reflects the signal strength within that frequency range, which is an important spectral feature. To simulate the nonlinear characteristics of human ear perception, the energy values of each band-pass filter are logarithmically transformed. The logarithmic transformation compresses the dynamic range, making the features more stable. Let the extracted English speech pronunciation feature be represented by σ_u, the center frequency of the filtered signal be represented by $ME(d)$, and the Fourier transform factor be represented by $d(a)$. The specific process is as follows:

$$\sigma_u = ME\left(d\right)d\left(a\right)/A_u \tag{13}$$

In the proposed method, the automatic correction of English pronunciation errors can be achieved through the confusion matrix factor clustering algorithm and Gaussian weighting method. The steps are detailed as follows:

Step 1: Obtain target English pronunciation factors

First, select several training data points from the learners' spoken pronunciation data. This training data should include various pronunciation scenarios, including correct pronunciations and common pronunciation errors. Next, use the confusion matrix to quantify the differences between the learners' pronunciations and standard English pronunciations. Each element in the confusion matrix represents the probability of a certain phoneme pronounced by the learner being recognized as another phoneme. To obtain the confusion probability values, an acoustic model can be used to perform forced alignment on the learners' pronunciations and the standard English pronunciations. Through forced alignment, we can acquire the target Hidden Markov sequence and preserve its temporal information.

Step 2: Identify correct English phonemes

After obtaining the confusion probabilities of the learners' pronunciation data, the next step is to identify the correct phonemes in English pronunciation. This step relies on a trained acoustic model that can effectively decode the learners' pronunciation data. The acoustic model analyzes the training data to recognize standard English phonemes and preserves the temporal information of these phonemes. Specifically, the acoustic model extracts feature from the input learners' pronunciation and decodes them into a phoneme sequence. The preserved temporal information includes the start and end times of each phoneme, which is crucial for subsequent co-occurrence matrix construction and error correction. By accurately preserving temporal information, we can align the correct phonemes with the learners' pronunciation data in time, providing data support for building the confusion matrix.

Step 3: Establish the co-occurrence matrix

After identifying the correct English phonemes and preserving their temporal information, we need to construct a related matrix, i.e., a co-occurrence matrix. Specifically, first, align the target Hidden Markov sequence obtained in Step 1 with the correctly identified phoneme sequence from Step 2 in time. Through alignment, we can determine the correspondence of each phoneme in the learners' pronunciation with the correct phonemes along the timeline. Based on the results of this time alignment, we construct the co-occurrence matrix. Each element in the co-occurrence matrix represents the co-occurrence frequency or probability between a certain phoneme in the learners' pronunciation and the correct phoneme. This matrix provides detailed distribution information about pronunciation errors, reflecting which phonemes are easily confused and the temporal distribution characteristics of this confusion. By analyzing the co-occurrence matrix, we can identify high-frequency pronunciation errors and provide specific temporal locations for each type of error. Let the co-occurrence matrix be represented by S, the number of correct consonant phonemes be represented by a, and the number of learner's English pronunciation phonemes be represented by b.

$$S = \left[a \times b\right] \tag{14}$$

Step 4: Similarity calculation

After constructing the co-occurrence matrix, we need to further quantify the similarity between the learners' pronunciations and standard English pronunciations. The key to this step is to determine the specific method for similarity calculation in order to form a similarity matrix. Let the number of correct English pronunciation phonemes be y and the number of learner pronunciation phonemes be o. We label the co-occurrence matrix as $S_{t.s}(y,o)$, which records the similarity between each correct phoneme and the learner pronunciation phonemes. Specifically, first represent the pronunciation features of each phoneme as vectors, where the features may include frequency, duration, energy, etc. Next, use the cosine similarity formula to calculate the similarity between each pair of phoneme vectors. Finally, fill in the similarity values for each pair of phonemes into the similarity matrix. Let the correct consonant pronunciation phoneme be represented by l_k,

the learner's consonant pronunciation phoneme by v_u, the similarity coefficient by *COUNT*, the number of learner consonant pronunciation phonemes by o, the number of tested pronunciation factors by O, and the similarity status matrix by $S(u,k)$.

$$S(u,k) = COUNT(l_k \mid v_u) / \sum_{o=1}^{o} COUNT(l_g, v_u) \quad (15)$$

Step 5: Information state mapping

After obtaining the similarity matrix, we need to perform state mapping for this similarity information. The goal of information state mapping is to extract the optimal and alternative correction states from the similarity matrix to guide the correction of pronunciation errors. Specifically, traverse the similarity matrix $S_{t.s}(y, o)$, and select the column with the highest similarity corresponding to each correct phoneme as the best correction state. Repeat this process to select several elements with high similarity, forming a series of correction states ac_v. Map these correction states into a high-dimensional space for better analysis and processing.

Step 6: Determine correction data

The final step is to calculate the correction coefficients for the learners' pronunciation based on the determined correction states and mapped data, and to complete the automatic correction. Specifically, based on the correction states mapped to high-dimensional space, calculate the correction coefficients for each correction state. The correction coefficients can be determined through linear regression or other optimization methods to ensure that the corrected pronunciation is as close as possible to the standard pronunciation. By combining the best correction states and alternative correction states, we comprehensively consider the weights of different correction states to form the final correction strategy. The weight allocation can be based on a weighted average of similarity values or optimized through methods such as Bayesian inference. Apply the correction coefficients to the learners' pronunciation data to adjust the feature parameters of each pronunciation phoneme, thus achieving automatic pronunciation correction. Specifically, parameters such as frequency, duration, and energy of the phonemes can be adjusted to make them closer to the standard pronunciation. Let the state correction weight be represented by α, the erroneous pronunciation data by $e_{u,ksg}(z_h)_{zq}$, the correction coefficient by e_{msg}, and the observation state by z_h.

$$e_{u,k^{sg}}(z_h) = \alpha e_{u,k^{sg}}(z_h) + 1 - \alpha \sum_{u=1}^{ac} \alpha_{uk} e_{m^{sg}}(z_h) \quad (16)$$

Through the above steps, we can systematically achieve automatic correction of learners' English pronunciation errors. The similarity calculation, information state mapping, and determination of correction data constitute a complete correction process, ensuring that learners can improve their pronunciation accuracy through scientific methods.

4. EXPERIMENTAL RESULTS AND ANALYSIS

From the experimental results in Table 1, it can be observed that different branch expansion rates have a significant impact on the WER in English speech recognition. When the expansion rate is set to "1-1-1-1-1-1-1-1," the WER is the highest at 9.11%, indicating that using the same expansion rate configuration fails to effectively capture the complex features of speech signals. As the expansion rate gradually increases and uses incremental configurations, the WER significantly decreases. For instance, with the expansion rate set to "1-3-6-9-11-13-17-21," the WER drops to 7.14%; when the expansion rate is set to "1-2-4-6-8-12-14-16," the WER reaches the lowest at 6.89%. This indicates that increasing differentiation in the expansion rate combinations helps to more accurately capture different hierarchical features in speech signals, thereby enhancing recognition accuracy. Based on the experimental results, it can be concluded that by designing a reasonable combination of expansion rates, the proposed speech recognition network significantly reduces the WER and demonstrates high recognition performance. Particularly, the expansion rate combination "1-2-4-6-8-12-14-16" effectively improves the network's ability to capture speech features, validating the advantages of this branch expansion rate configuration in speech recognition tasks.

Table 1. Impact of different branch expansion rates on word error rate (WER) in English speech recognition

Expansion Rate	WER (%)
1-1-1-1-1-1-1-1	9.11
1-1-1-1-2-3-7-9	7.62
1-3-6-9-11-13-17-21	7.14
1-2-3-4-5-6-7-8	7.32
1-2-4-6-8-12-14-16	6.89

Table 2. Impact of number of branches on WER in English speech recognition

Number of Branches	Parameter Count (M)	WER (%)	Training Duration (hours)
4	12.9	7.13	≈37
8	17.8	6.88	≈55
15	27.6	6.49	≈82

Table 3. Impact of different numbers of attention heads on WER in English speech recognition

Use of Deep Convolutional Neural Network	Number of Attention Heads	Attention Subspace Dimension	WER (%)
No	4	63	9.11
	8	31	9.24
Yes	2	125	8.7
	4	62	6.89
	8	31	6.55

From Table 2, it is clear that increasing the number of branches has a noticeable optimizing effect on the WER in English speech recognition. When the number of branches is 4, the model's WER is 7.13%, with a parameter count of 12.9M and a training duration of approximately 37 hours. As the number of branches increases to 8, the WER decreases to 6.88%, with the parameter count rising to 17.8M and the training duration increasing to about 55 hours. When the number of branches is further increased to 15, the WER significantly drops to 6.49%, with a parameter count reaching 27.6M and a training duration of about 82 hours. The data indicates that while increasing the number of branches raises the model's complexity and training duration, it also

significantly enhances the model's speech recognition accuracy. It can be concluded that the proposed multi-branch structured English speech recognition network is clearly effective in optimizing the WER. Increasing the number of branches allows the model to better capture and process subtle features in speech signals, improving recognition accuracy; although the parameter count and training duration increase, this performance cost is justified by the significant reduction in WER.

From the experimental results in Table 3, it can be seen that the number of attention heads and the use of Deep Convolutional Neural Networks (DCNN) significantly affect the WER in English speech recognition. In the absence of DCNN, the WER for 4 and 8 attention heads are 9.11% and 9.24%, respectively, with a corresponding decrease in subspace dimension (from 63 to 31), indicating that the attention mechanism has limited improvement on the WER without the use of deep convolution. However, when using DCNN, the WER drops significantly, especially with 4 and 8 attention heads, where the rates fall to 6.89% and 6.55%, respectively. Under the combined influence of DCNN and multi-head attention, the model significantly improves speech recognition accuracy. Based on the above experimental data, it can be concluded that combining DCNN with a multi-head attention mechanism is an effective strategy for enhancing speech recognition accuracy. In the absence of DCNN, increasing the number of attention heads does not significantly improve the WER; however, with the support of DCNN, the WER decreases significantly, and the enhancing effect of the multi-head attention mechanism becomes more pronounced. The results validate that the network structure proposed in this study can effectively enhance speech recognition performance, particularly in improving recognition accuracy, demonstrating superior performance and a rational technical pathway for practical applications.

Table 4. Impact of different numbers of convolution blocks and internal convolution kernel sizes on WER in English speech recognition

Number of Convolution Blocks	Expansion Factor	Parameter Count (M)	WER (%)
4	2-2-2-2	16.8	7.23
8	3-3-...-3	17.9	6.58
8	1-2-2-2-4-4-4-1	21.3	6.97
8	1-3-3-6-6-3-3-1	18.8	6.65
15	2-2-...-2	21.2	6.79

Table 5. Ablation experiment of SENet and feedforward neural network

	Last Layer Dimension	SENet & Feedforward Neural Network Dimension	WER (%)
-(*SENet*+ Feedforward)	524	(0,0)	8.88
-*SENet* -Feedforward	524	(0,524)	7.23
	524	(524,0)	7.62
SENet+ Feedforward	524	(524,524)	6.89
-(*SENet*+ Feedforward)	524	(0,0)	11.24
SENet+ Feedforward	524	(262,262)	9.84

From the experimental results in Table 4, it is evident that the number of convolution blocks and the internal convolution kernel sizes significantly affect the WER in English speech recognition. When the number of convolution blocks is 4 and the expansion factor is "2-2-2-2," the WER is 7.23% with a parameter count of 16.8M. As the number of convolution blocks increases to 8 with an expansion factor of "3-3-...-3," the WER decreases to 6.58%, while the parameter count slightly increases to 17.9M. With 8 convolution blocks and an expansion factor configuration of "1-2-2-2-4-4-4-1," the parameter count rises to 21.3M, but the WER is 6.97%, indicating a delicate balance between parameter count and WER. When the number of convolution blocks is 15 with an expansion factor of "2-2-...-2," the WER is 6.79%, and the parameter count increases to 21.2M, suggesting that more convolution blocks can enhance recognition accuracy, though the effectiveness depends on the specific configuration. Based on the experimental results, it can be concluded that the reasonable configuration of the number of convolution blocks and the expansion factors is a crucial factor for improving speech recognition accuracy. Increasing the number of convolution blocks and selecting appropriate expansion factor combinations can better capture and represent the details of speech signals, as demonstrated by the significant reduction in WER to 6.58% with the "3-3-...-3" configuration of 8 convolution blocks. These experimental results validate the effectiveness of the proposed English speech recognition network based on convolution blocks and expansion factor optimization, highlighting the significant impact of optimizing convolution configurations within a certain parameter range on enhancing speech recognition performance.

From the ablation experiment results in Table 5, it can be observed that the SENet module and the Feedforward Neural Network (FFN) each contribute to improving speech recognition accuracy. When only the SENet module is used (dimension configuration (0,0)), the WER is 8.88%. In contrast, when only the FFN module is used (dimension configuration (0,524)), the WER decreases to 7.23%, indicating that the FFN module has a significant role in enhancing recognition accuracy. When only the SENet dimension is set to 524, the WER is 7.62%. Furthermore, when combining SENet and FFN (dimension configuration (524,524)), the WER significantly drops to 6.89%, demonstrating the collaborative effect of both modules in improving speech recognition performance. In comparison, a lower dimension combination of SENet and FFN (262,262) yields a WER of 9.84%, indicating that reducing the dimensions diminishes the optimization effect of both modules. Based on the above experimental results, it can be concluded that the combination of SENet and FFN modules can effectively reduce the WER and improve speech recognition accuracy. While using SENet or FFN alone does yield some improvement, their synergistic effect significantly enhances model performance, especially at higher configuration dimensions, achieving the lowest WER of 6.89%. This optimization effect validates the effectiveness of the network structure proposed in this study, indicating that enhancing the network's adaptability and representational capability for features can significantly improve recognition accuracy, providing important technical support for achieving high-precision English speech recognition.

Figure 4 shows that the proposed method outperforms DTW-CNN and Wav2Vec in total number of corrections across different testing instances. At 50 tests, the total number

of corrections for the proposed method is 275, exceeding DTW-CNN's 265 and Wav2Vec's 240. As the number of tests increases, the total corrections for the proposed method gradually decline but remain superior throughout the testing process. By 300 tests, the proposed method's corrections total 210, significantly higher than DTW-CNN's 160 and Wav2Vec's 185. This indicates that the proposed method demonstrates strong stability and durability under various testing conditions, particularly at higher test counts where its advantages are most pronounced. Based on the experimental results, it can be concluded that the proposed signal processing-based English speech pronunciation correction method exhibits clear superiority in the number of corrections. Compared to DTW-CNN and Wav2Vec, the proposed method maintains a high number of corrections across multiple test instances, highlighting its advantages in accuracy and durability.

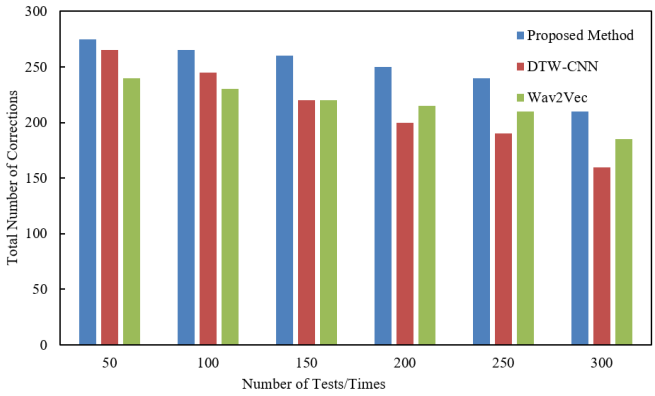

Figure 4. Total number of corrections for different methods

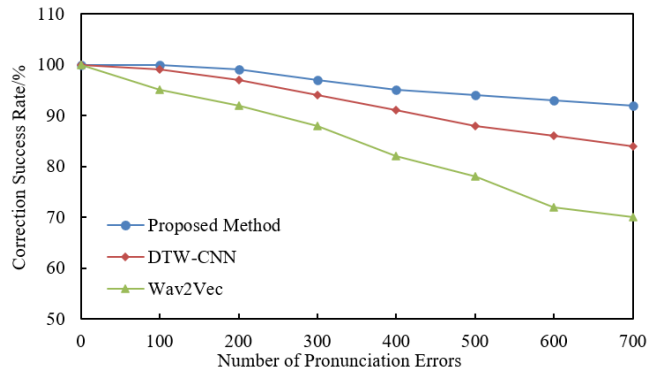

Figure 5. Correction success rate test results for different methods

The results in Figure 5 show that the proposed method consistently outperforms DTW-CNN and Wav2Vec in terms of correction success rate across varying numbers of pronunciation errors. When the number of pronunciation errors is 0, all three methods achieve a 100% correction success rate. However, as the error count increases, the differences become more pronounced. Specifically, when the number of pronunciation errors reaches 300, the success rate of the proposed method remains at 97%, while DTW-CNN and Wav2Vec drop to 94% and 88%, respectively. When the error count increases to 700, the proposed method maintains a correction success rate of 92%, significantly higher than DTW-CNN's 84% and Wav2Vec's 70%. This demonstrates that the proposed method retains a high correction success rate even in the presence of numerous errors. Based on the experimental data, it can be concluded that the proposed signal processing-based English pronunciation correction method exhibits significant advantages in correction success rate, especially when the number of pronunciation errors is high. Compared to DTW-CNN and Wav2Vec, the proposed method demonstrates strong resistance to interference and stability, ensuring high-quality correction outcomes in complex pronunciation error environments.

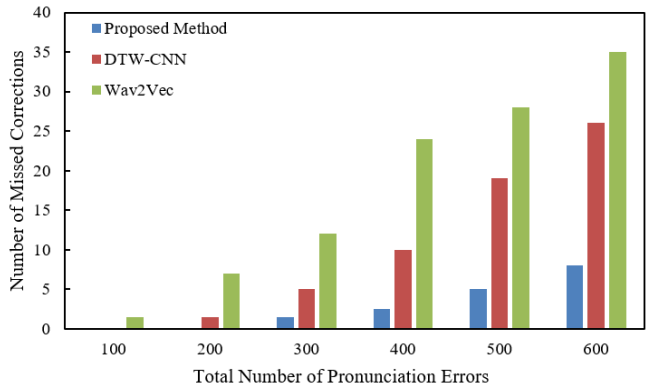

Figure 6. Missed correction count test results for different methods

The test results for missed correction counts in Figure 6 indicate that the proposed method consistently achieves a lower count of missed corrections compared to DTW-CNN and Wav2Vec across varying total numbers of pronunciation errors, demonstrating clear advantages. When the total number of pronunciation errors is 100, the missed correction counts for the proposed method, DTW-CNN, and Wav2Vec are 0, 0, and 1.5, respectively, indicating effective pronunciation correction by all methods under low error conditions. However, when the total number of errors increases to 300, the proposed method's missed correction count is only 1.5, while DTW-CNN and Wav2Vec's counts rise to 5 and 12, respectively. At a total error count of 600, the proposed method's missed correction count is 8, while DTW-CNN and Wav2Vec report 26 and 35, respectively, showing that the proposed method maintains a low missed correction count even with many pronunciation errors. Based on the experimental data, it can be concluded that the proposed signal processing-based English pronunciation correction method demonstrates significant advantages in terms of missed correction counts, especially when faced with a high number of pronunciation errors. Compared to DTW-CNN and Wav2Vec, the proposed method not only maintains good correction performance under low error conditions but also exhibits strong robustness and resistance to interference under high error conditions. These results validate the effectiveness of the proposed method, indicating that it can provide high-quality correction support in complex pronunciation correction environments, ensuring learners receive more accurate pronunciation correction experiences.

5. CONCLUSION

This study developed an English speech recognition network and pronunciation correction method based on signal processing technology, aiming to enhance recognition accuracy and correction precision, thereby providing personalized pronunciation guidance for learners. The

research covers the impact of various factors, including different expansion rates, branch numbers, attention head counts, convolution block counts, and kernel sizes on the WER in speech recognition. Experimental results show that reasonably designed combinations of expansion rates and branch structures can effectively reduce the WER, particularly through the use of varying attention head counts and optimized convolution structures, which enhance the network's feature extraction capability and further improve recognition accuracy. Moreover, ablation experiments reveal that the combination of SENet and feedforward neural networks significantly enhances model performance, outperforming traditional methods like DTW-CNN and Wav2Vec in terms of correction success rate and missed correction count, validating the robustness of the proposed methods in multiple scenarios.

In summary, the methodologies employed in this research significantly can improve the effectiveness of English speech recognition and pronunciation correction, showcasing the strong capabilities of signal processing-based models in addressing complex speech features. However, the study also identifies limitations, such as the model's computational intensity, which may affect real-time performance, and the need for further optimization regarding adaptability to specific noisy environments. Future research directions could focus on enhancing the model's computational efficiency, optimizing robustness against various background noises, and exploring adaptive speech feature enhancement techniques to further improve the practicality of recognition and correction. This research provides new methods and insights for the fields of speech recognition and pronunciation correction, laying the groundwork for more intelligent and efficient language learning support systems.

REFERENCES

[1] Qin, X. (2023). Research on auxiliary training system of oral English pronunciation based on data extraction. SN Applied Sciences, 5(3): 84. https://doi.org/10.1007/s42452-023-05306-x

[2] Zhang, F. (2023). Research on automatic annotation of English pronunciation errors based on deep transfer learning. International Journal of Computer Applications in Technology, 73(4): 245-252. https://doi.org/10.1504/IJCAT.2023.138828

[3] Nguyen, L.T., Burri, M. (2024). Pronunciation pedagogy in English as a foreign language teacher education program in Vietnam. International Review of Applied Linguistics in Language Teaching, 62(2): 675-691. https://doi.org/10.1515/iral-2022-0126

[4] Liu, Y., Quan, Q. (2022). AI recognition method of pronunciation errors in oral English speech with the help of big data for personalized learning. Journal of Information & Knowledge Management, 21(2): 2240028. https://doi.org/10.1142/S0219649222400287

[5] Sheng, Y., Yang, K. (2021). Automatic correction system design for English pronunciation errors assisted by high-sensitivity acoustic wave sensors. Journal of Sensors, 2021(1): 2853056. https://doi.org/10.1155/2021/2853056

[6] Abad-Célleri, M., Argudo-Serrano, J., Fajardo-Dack, T., Cabrera, P. (2024). Ecuadorian EFL preservice teachers' attitudes toward pronunciation features. Profile Issues in Teachers Professional Development, 26(1): 81-96.

[7] Xiao, W., Park, M. (2021). Using automatic speech recognition to facilitate English pronunciation assessment and learning in an EFL context: Pronunciation error diagnosis and pedagogical implications. International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT), 11(3): 74-91. https://doi.org/10.4018/IJCALLT.2021070105

[8] Zhang, S.F. (2021). Design of an automatic English pronunciation error correction system based on radio magnetic pronunciation recording devices. Journal of Sensors, 2021(1): 5946228. https://doi.org/10.1155/2021/5946228

[9] Xue, X.J., Dunham, R.E. (2023). Using a SPOC-based flipped classroom instructional mode to teach English pronunciation. Computer Assisted Language Learning, 36(7): 1309-1337. https://doi.org/10.1080/09588221.2021.1980404

[10] Wang, S., Shi, X. (2021). Research on correction method of spoken pronunciation accuracy of AI virtual English reading. Advances in Multimedia, 2021(1): 6783205. https://doi.org/10.1155/2021/6783205

[11] Uchida, Y., Sugimoto, J. (2020). Non-native English teachers' confidence in their own pronunciation and attitudes towards teaching: A questionnaire survey in Japan. International Journal of Applied Linguistics, 30(1): 19-34. https://doi.org/10.1111/ijal.12253

[12] He, H. (2022). Design of a speaking training system for English speech education using speech recognition technology. International Journal of Advanced Computer Science and Applications, 13(11): 450-455. https://doi.org/10.14569/IJACSA.2022.0131151

[13] Mahdi, H.S., Alkhammash, R., Al-Athwary, A.A. (2024). Using high variability phonetic training as a contextualized tool in the development of English consonant clusters pronunciation among Saudi EFL learners. Education and Information Technologies, 29(6): 6821-6840. https://doi.org/10.1007/s10639-023-12113-9

[14] Shanmugasundaram, R., Jebakumar, A.N. (2022). Mother tongue influence on English pronunciation: A case study in college students. Journal for Educators, Teachers and Trainers, 13(4): 312-316. https://doi.org/10.47750/jett.2022.13.04.042

[15] Alghonaim, A.S. (2020). Impact of watching cartoons on pronunciation of a child in an EFL setting: A comparative study with problematic sounds of EFL learners. Arab World English Journal, 11(1): 52-68. https://doi.org/10.24093/awej/vol11no1.5

[16] Indrayadi, T., Daflizar, D., Helty, H. (2021). Indonesian EFL students' difficulties in recognizing English letters. Qualitative Report, 26(11): 3476-3491. https://doi.org/10.46743/2160-3715/ 2021.4846

[17] Phuong, T.T.H. (2021). Who should teach English pronunciation? Voices of Vietnamese EFL learners and teachers. Journal of Asia TEFL, 18(1): 125-141.

[18] Cao, H., Dong, C. (2022). An English pronunciation error detection system based on improved random forest. Mobile Information Systems, 2022(1): 6457286. https://doi.org/10.1155/2022/6457286

[19] Rehman, I., Silpachai, A., Levis, J., Zhao, G., Gutierrez-Osuna, R. (2022). The English pronunciation of Arabic speakers: A data-driven approach to segmental error identification. Language Teaching Research, 26(6): 1055-1081. https://doi.org/10.1177/1362168820931888

[20] Alghazo, S., Jarrah, M., Al Salem, M.N. (2023). The

efficacy of the type of instruction on second language pronunciation acquisition. Frontiers in Education, 8: 1182285. https://doi.org/10.3389/feduc.2023.1182285

[21] Tsunemoto, A., McDonough, K. (2021). Exploring Japanese EFL learners' attitudes toward English pronunciation and its relationship to perceived accentedness. Language and Speech, 64(1): 24-34. https://doi.org/10.1177/0023830919900372

[22] Wang, Y., Zhao, P. (2020). A probe into spoken English recognition in English education based on computer-aided comprehensive analysis. International Journal of Emerging Technologies in Learning (iJET): 15(3): 223-233.

[23] Havlik, M. (2020). The pronunciation of dental fricatives in Anglicisms and English proper names in Czech: codification, norms and use. Slovo a Slovesnost, 81(3): 195-217.