Hindawi
Journal of Sensors
Volume 2022, Article ID 7426303, 13 pages
https://doi.org/10.1155/2022/7426303

Research Article

Oral Business English Recognition Method Based on RankNet Model and Endpoint Detection Algorithm

Xia Xu[1] and Kunxue Xiao[2]

[1]*School of Foreign Languages, Xinyang University, Xinyang, Henan 464000, China*
[2]*School of Foreign Languages and Literatures, Chongqing University of Education, Chongqing 400065, China*

Correspondence should be addressed to Kunxue Xiao; xiaokx@cque.edu.cn

Received 20 March 2022; Revised 31 March 2022; Accepted 4 April 2022; Published 11 May 2022

Academic Editor: Yuan Li

The prosodic feature of "stressed syllables" plays an important role in business English learning. For the majority of nonnative language learners learning English, the pronunciation of native language can easily affect the expression of spoken English. For those who cooperate and communicate in business, if they want to have more idiomatic and accurate spoken English, they need to control the necessary condition of stress first. This puts forward new requirements for the existing speech recognition technology: more accurate recognition of stressed syllables, reading complex and different emotional colors, and helping to correct oral expression. The experimental results show that (1) the higher the Fisher Ratio of the feature, the easier it is to distinguish the stressed syllables; adjusting the weight of features can effectively improve the recognition accuracy. (2) The recognition rate will decrease with the increase of noise. (3) InwMS method can distinguish features better than min-max method, but the recognition rate of stressed syllables is very low. Linear recognition based on single feature is not recommended. (4) The error rates of the two methods are 20.6% and 19.32%. If any feature of the fusion feature is removed, the error recognition rate of the model will increase by at least 3%. (5) For sentence stress recognition based on fusion features, the recognition error rate on RankNet model is as high as 42.51%. The final result of system operation is good, simple, and convenient.

1. Introduction

With the rapid development of economic globalization, neither the country nor the individual can stay out of it, and the whole world is closely linked. If people want to seek better cooperation, mutual communication and help are essential and important factors. Combined with the rapid development of computer technology in recent decades, using intelligent systems, software, and other methods to assist in learning business English has been widely popular in people's study, work, and life. Due to the wide circulation of English, it occupies an important position in economic and trade exchanges. How to learn oral English well and use it in business cooperation is a hot topic that occupies the study list for a long time. In this paper, researchers through the study of a lot of relevant literature summed up the previous experiments to get some empirical conclusions and make use of the discovery to create new value.

Compare the syllables and morphemes of simple words and compound words [1]. It is found that lexical parameters and audibility have influence on word speech recognition in monosyllabic speech test [2]. The quality of final vowels of words plays an important role in English major stress [3]. Acoustic cues of word stress promote offline and online language processing to a certain extent [4], because speakers are sensitive to auditory feedback in stressed and unstressed syllables [5]. The effects of stress are analyzed for syllable position, pitch, duration, and amplitude [6]. Using four deep neural network architectures, the problems of phonetic symbols, lexical stress assignment, and syllability can be solved [7]. In order to improve the recognition ability of accented syllables in spoken English, a word stress recognition model is established based on natural language processing and endpoint detection algorithm [8]. According to syllable stress, transfer learning and CNN neural network are used to recognize English phonetic emotion [9]. When the speech cluster is in

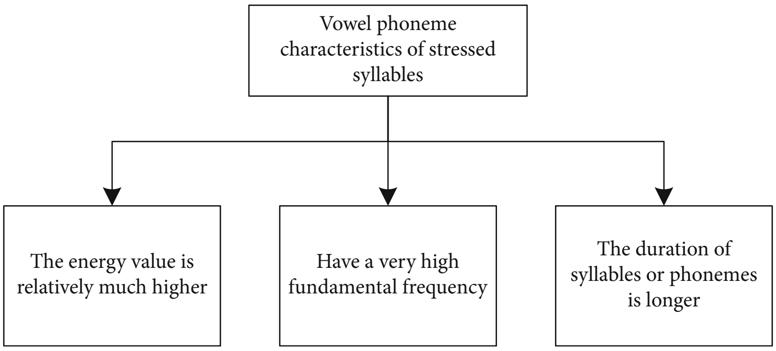
Vowel phoneme characteristics of stressed syllables
The energy value is relatively much higher
Have a very high fundamental frequency
The duration of syllables or phonemes is longer

Figure 1: Factor characteristics of stress.

the middle of the word, the stress has a great influence, and the recognition is more accurate [10]. If you realize stress placement and emphasize syllables when learning English, it is helpful to learn English words [11]. The hybrid algorithm based on HMM and clustering is used to recognize English speech, which improves the recognition speed by 20.1% [12]. After experiments under different parameters, the CNN algorithm optimized by LPC has higher accuracy of English speech recognition [13]. CTC word model for acoustic detection of speech recognition needs to train more data and has better performance [14]. Applying speech recognition to college English listening teaching can effectively improve students' listening [15].

The above documents are explained from different methods and angles. It not only shows the importance and related characteristics of learning oral stress, but also provides us with excellent ideas and good experimental methods for our research. Although there are many speech research results that are good, but so far the speech system targeted is not strong, and speech recognition technology is not perfect and mature. In addition to the inability to accurately identify the position of learners' stressed pronunciation, learners cannot get a reasonable assessment report of stressed pronunciation, which cannot effectively achieve the purpose of oral business English learning. There is not yet a successful business English learning system for recognizing stressed syllables in spoken English. Most systems are faced with the problems of incomplete recognition function and imperfect performance. Therefore, this paper uses natural language processing and endpoint detection algorithm to integrate the technology of the system, so as to have a better application in speech recognition of stressed syllables.

2. Theoretical Basis

2.1. Oral Business English. Business English [16], as the name implies, mainly involves business interaction. Under the trend of internationalization, the purpose of this type of language is to help learners adapt to work and life in the workplace. In addition to simple and basic English language learning, learners not only need to improve their English expression level, but also need to understand business terminology, western business management concepts, and conventions; Foreigners' thinking mode and psychological state are different from our own country, which involves the conceptual requirements at the cultural level. If you want to do business with foreigners well, you need to know the ways and means to cooperate with them and respect their business requirements and standards.

Therefore, the functionality of "spoken language" is very important, which has three main components: words, grammar, and sentences. Oral fluency and pronunciation accuracy are very important in real interpersonal communication. As it happens, English is a language of stress and rhythm. It is self-evident that stress plays an important role in learning words and sentences, which helps to lay a solid foundation for rhythm and intonation, and it includes important grammatical and semantic characteristics. Some literature also well illustrates that the core of mastering business English and oral English learning is the stressed syllables of words and sentences. It is shown Figure 1:

2.2. Natural Language Processing. Natural language processing [17] (NLP) is an interdisciplinary subject of intelligence science and English. For machines, digital information is its expression. Through the bridge of NLP, we can create a way for human beings to communicate with machines, so that human beings can understand what machines are communicating and expressing. However, NLP also has five major difficulties: (1) language has no simple and intuitive rules. Its laws are complicated, and it is difficult for ordinary people to understand and observe. (2) Openness of English can be interconnected with other languages. (3) Language is an open collection. Users can discover and create new expressions by themselves. (4) Language is often associated with concrete practice, and it depends on the knowledge of various disciplines. (5) If you want to use a language well and smoothly, you need to base yourself on the environment and context of the language, as shown in Table 1:

2.3. Endpoint Detection Algorithm. Speech signal processing is very important. Endpoint detection algorithm technology is used to distinguish speech and nonspeech signals, so as to accurately determine the starting point and ending point of speech. In reality, the real environment of speech recognition is easily affected by noise, so the endpoint detection algorithm should reduce the interference of mute and noise. If the detection of speech signal endpoint is not accurate, it is easy to cause recognition errors, which directly determines the success or failure of the whole speech recognition work in some aspects. The performance of speech recognition can be

Table 1: Four applications of NLP.

Typical applications of NLP	
Emotional analysis [18]	The network is full of complex and redundant text information. These messages have different effects for different people or machines. However, if you make use of these texts. It can be found that the emotions they express are nothing more than positive or negative feedback
Machine translation	Language and language need to be translated before undifferentiated communication can be realized. Nowadays, the application of machine translation is mature, such as Google translation and Baidu translation
Chat robot [19]	It represents the future intelligent application, which promotes the application of chat robots to produce more new values. Using this kind of robot, smart home, intelligent entertainment, and other aspects glow with new vitality
Speech recognition [20]	The application of speech recognition has entered thousands of households and is widely used in medical treatment, communication, home appliances, games, and other industries. Whether it is the voice-to-text function in QQ or WeChat, voice navigation for cars etc., its use is not a complex laboratory high-tech, but a national-level application

improved effectively if the collection amount of speech data is effectively processed and the processing time is saved.

(1) Short-term average energy [21]

$$E_n = \sum_{m \longrightarrow -\infty}^{\infty} [x(m) \cdot \omega(n-m)]^2. \tag{1}$$

For $h(n) = \omega^2(n)$, there are

$$E_n = \sum_{m \longrightarrow -\infty}^{\infty} x^2(m) \cdot h(n-m). \tag{2}$$

Short-term average amplitude:

$$M_n = \sum_{m \longrightarrow -\infty}^{\infty} |x(m)| \cdot \omega(n-m), \tag{3}$$

where $\omega(n)$ is Hamming window and $h(n)$ can realize framing processing. When the value obtained by $E(n)$ is large, it can be applied to the voiced segment, and when the value obtained by $E(n)$ is small, it can be applied to the unvoiced segment, distinguishing the starting point and the ending point of the speech signal. E_n and M_n reflect the strength of the speech signal.

(2) Short-term average zero-crossing rate

$$Z_n = \sum_{m \longrightarrow -\infty}^{\infty} |\operatorname{sgn}[x(n)] - \operatorname{sgn}[x(n-1)]| \cdot \omega(n-m), \\ \operatorname{sgn}[x(n)] = \begin{cases} 1, x(n) \geq 0, \\ -1, x(n) < 0. \end{cases} \tag{4}$$

General take:

$$\omega(n) = \begin{cases} \dfrac{1}{2N}, & 0 \leq n \leq N-1, \\ 0, & \mathit{else}. \end{cases} \tag{5}$$

Threshold zero crossing rate [22]:

$$Z_n = \frac{1}{2} \sum_{m \longrightarrow -\infty}^{\infty} \{|\operatorname{sgn}[x(n) - T] - \operatorname{sgn}[x(n-1) - T]| + |\operatorname{sgn}[x(n) + T] - \operatorname{sgn}[x(n-1) + T]|\} \cdot \omega(n-m). \tag{6}$$

2.4. Speech Feature Extraction

(1) Energy characteristics

$$Q_n = T[x(m)w(n-m)]. \tag{7}$$

(2) Duration characteristics. Longer syllables can be judged as stressed syllables

(3) Fundamental frequency characteristics [23]

Discrete signal sequence:

$$R(k) = \sum_{m \longrightarrow -\infty}^{\infty} x(m) \cdot x(m+k). \tag{8}$$

Random signal sequence:

$$R(k) = \lim_{N \longrightarrow -\infty} \frac{1}{2N+1} \sum_{m \longrightarrow -N}^{\infty} x(m) \cdot x(m+k). \tag{9}$$

Autocorrelation function:

$$\begin{aligned} x(k) &= x(n+N_P), \\ R(k) &= R(k+N_P). \end{aligned} \tag{10}$$

(4) LPC [24]

$$\begin{aligned} R(i) &= \sum_{n=-\infty}^{+\infty} S_w(n+i) S_w(n), \\ H(z) &= S(z)/U(z) = G/\left(1 - \sum_{k=1}^{P} \alpha_k z^{-k}\right), \\ S(z) &= \sum_{k=1}^{p} \alpha_k S(n-k) + Gu(n). \end{aligned} \tag{11}$$

(5) Mel cepstrum coefficients

Output of triangular filter:

$$Y_i = \sum_{k=F_{i-1}}^{F_i} \frac{k - F_{i-1}}{F_i - F_{i-1}} X_k + \sum_{k=F_{i+1}}^{F_{i+1}} \frac{F_{i-1} - k}{F_{i+1} - F_i} X_k, i + 1, 2, \cdots, 24. \tag{12}$$

DCT Transform [25]:

$$C_k = \sum_{j=1}^{24_i} \log\left(Y_j\right) \cos\left[k\left(j - \frac{1}{2}\right)\frac{\pi}{24}\right], k = 1, 2, \cdots, p. \tag{13}$$

First-order difference cepstrum calculation:

$$\Delta c_1(m) = \sum_{k=-2}^{2} k c_{l-k}(m), 1 \le m \le p. \tag{14}$$

Mel frequency conversion relation:

$$Mel = \ln\left(1 + \frac{f}{700}\right) \cdot \frac{1000}{\ln\left(1 + 1000/700\right)}. \tag{15}$$

(6) Fractal dimension feature extraction

Box counting dimension:

$$D_B(s) = \lim_{\varepsilon \longrightarrow 0} \frac{\ln\left(N(\varepsilon)\right)}{\ln\left(1/\varepsilon\right)}. \tag{16}$$

Minkowski dimension:

$$\begin{aligned} D_M &= \lim_{\varepsilon \longrightarrow 0}\left(2 - \frac{\ln\left(A_G(\varepsilon)\right)}{\ln\left(\varepsilon\right)}\right), \\ A_G(\varepsilon) &= area\left(\underset{t \in F}{\cup} G_\varepsilon(t, f(t))\right). \end{aligned} \tag{17}$$

2.5. Algorithm Evaluation Index

(1) Recognition of stressed syllables of words

$$\begin{aligned} R_{\text{right}} &= N_{\text{right}}/\left(N_{\text{right}} + N_{\text{error}}\right), \\ R_{\text{error}} &= N_{\text{error}}/\left(N_{\text{right}} + N_{\text{error}}\right), \\ R_{\text{miss}} &= N_{\text{miss}}/\left(N_{\text{right}} + N_{\text{error}} + N_{\text{miss}}\right). \end{aligned} \tag{18}$$

(2) Recognition of sentence stress

$$R_i = \frac{N_i}{N}. \tag{19}$$

Among them, R_{right}, R_{error}, R_{miss}, R_i represents the correct recognition rate, wrong recognition rate, omission rate, and rereading recognition accuracy of evaluation indicators, respectively.

3. Recognition of Stressed Syllables Based on RankNet

3.1. Speech Recognition Process

3.1.1. NLP Corpus Preprocessing. The original speech data of NLP needs to be preprocessed. Mark the grammatical categories of words in sentences. Finally, the proper nouns in the text are identified and divided into blocks. First of all, the English is preprocessed to achieve the desired effect, and then after data cleaning. The text is segmented by using the word segmentation tool. Then the main part of the word is extracted, which is a process of getting the root after removing the prefix of the word. Word form is to use dictionaries to restore complex words to the most basic forms. Then, the grammatical categories of words in sentences are marked. Finally, the proper nouns in the text are identified and divided into blocks, as shown in Figure 2:

3.1.2. Stress Recognition Process. Compared with the traditional stressed syllable recognition process, it is relatively simple, and its functions are not perfect. In this paper, English words and sentences should be stressed and recognized. Therefore, we specially set up a two-layer stress recognition process, as shown in Figures 3 and 4:

The first layer mainly tests words. After feature extraction, phonemes are aligned according to time, and phoneme corpus is labeled to form training sample data. The second layer is sentence stress recognition. When the first layer

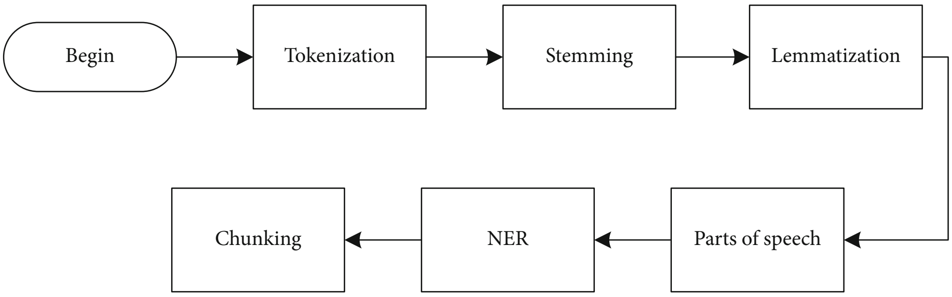

Figure 2: Steps of English corpus preprocessing.

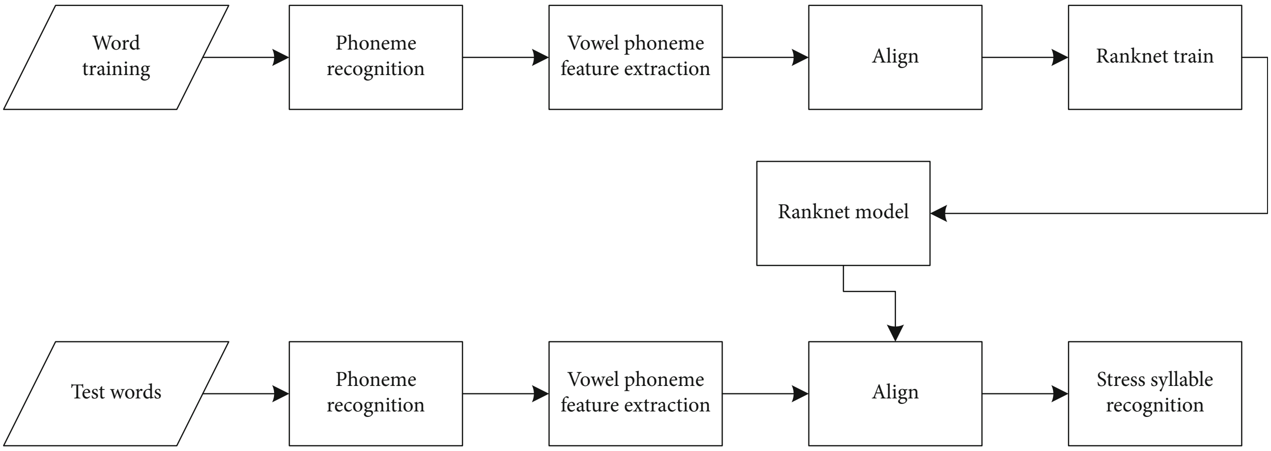

Figure 3: Layer 1 process: word stress recognition.

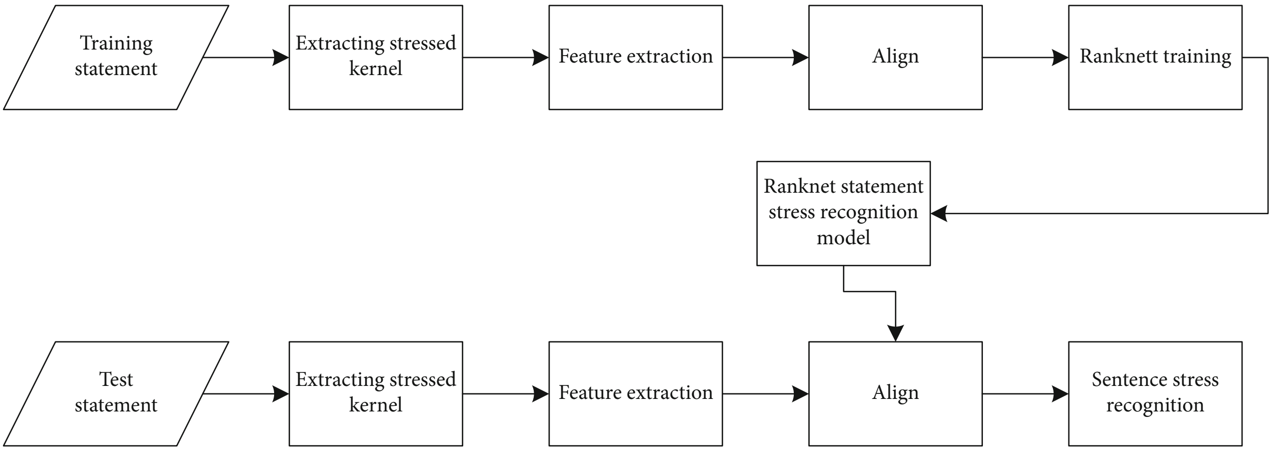
Training statement
Extracting stressed kernel
Feature extraction
Align
Ranknett training
Ranknet statement stress recognition model
Test statement
Extracting stressed kernel
Feature extraction
Align
Sentence stress recognition

Figure 4: Layer 2 process: statement stress recognition.

recognizes the word with each stressed syllable, the stressed syllable kernel of the word is taken out to participate in the recognition of sentence stress.

3.2. RankNet Model. RankNet is based on artificial neural network. It mainly solves the ranking from the perspective of probability. This self-learning technique forms a pair of data functions. Its basic processing unit is neuron. It consists of an output y_i and a large number of input objects x_i:

$$Y_j(t) = f\left(\sum_{i=1}^{n} w_{ji}x_i - \theta_j\right). \tag{20}$$

RankNet's network model is shown in Figure 5:

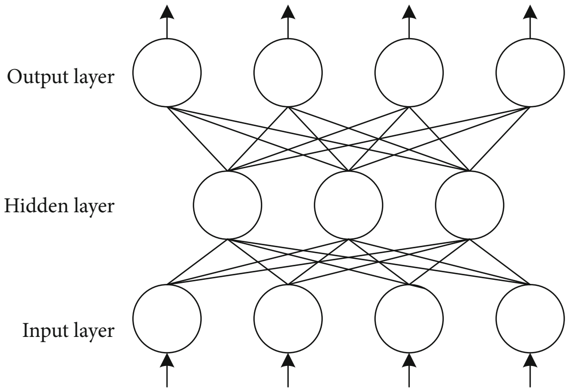

Figure 5: RankNet network architecture.

Table 2: RankNet.

RankNet	BP neural network
The input layer of RankNet is mainly composed of pairs of feature vectors	The input layer of typical BP neural network is very different, that is, feature vectors one by one
The loss function is mainly the output cross entropy of paired inputs. In back propagation, the output cross entropy loss function of a pair of input eigenvectors is used. According to the gradient of the connection weights of each layer, the connection weights between the input layer and the hidden layer and between the hidden layer and the output layer are adjusted	For BP neural network, its typical loss function is minimum mean square error. The interlayer connection weight is updated according to the error value between the actual output and the expected output

RankNet uses gradient descent algorithm and uses iterative operation to add hidden nodes to increase the adjustable parameters, thus obtaining the approximate exact solution, as shown in Table 2:

3.3. Speech Endpoint Detection. Speech endpoint detection is mainly used in silence, transition, sound and end stages. At different stages, their state transitions are different, where E H, EL, represent the high and low thresholds value of the short-term energy; RH, RL represent the high and low thresholds of the autocorrelation function value. As shown in Figure 6:

One threshold is relatively low and sensitive to signal changes; one threshold is higher. The dual-threshold endpoint detection algorithm is mainly based on short-term energy and zero-crossing rate fusion algorithm. Two different thresholds can be determined for speech endpoint detection by these two algorithms. One threshold is relatively low and sensitive to signal changes; one threshold is relatively high; when the double threshold is exceeded for a period of time, the two are properly matched, which means that the starting endpoint of the speech signal can be judged.

$$\begin{gathered} ZT = \min\left(\mathrm{IF}, zc + 2fzc\right), \\ ETL = \min\left(I_1, I_2\right), \\ ETU = 5ETL, \\ I_1 = 0.03(EMAX - EMIN) + \mathrm{EMIN}, \\ I_2 = 4EMIN. \end{gathered} \tag{21}$$

Among them, IF is generally 25, which represents the empirical value; zc and fzc represent the mean and standard deviation of the zero-crossing rate, respectively. The short-term average energy or average amplitude of each frame is recorded as E; the maximum value and the minimum value are recorded as $EMAX$ and $EMIN$, respectively.

DTW algorithm can save computing time, as shown in Figure 7:

3.4. Stress Recognition System. In this chapter, we mainly explain the overall framework of stressed syllables in the system. The final system structure diagram includes three sections. The first section is the feature extraction module, which mainly provides preliminary judgment and evaluation results. This module is the data basis of stressed syllables. The second section is a single feature analysis module, in which stressed syllables are judged by linear discriminant function; fractal dimension curve marks position. Finally, the location information is sent to the final section of the speech evaluation. This section can specifically train and analyze the evaluation results of statistics, as shown in Figure 8:

4. Simulation Experiment Analysis

After a series of preliminary design, the specific process and implementation of each module has been described. Our accented syllable recognition system for oral business English is also basically completed.

4.1. Feature Selection and Evaluation. The selection of speech features needs a certain judgment basis. We can optimize the

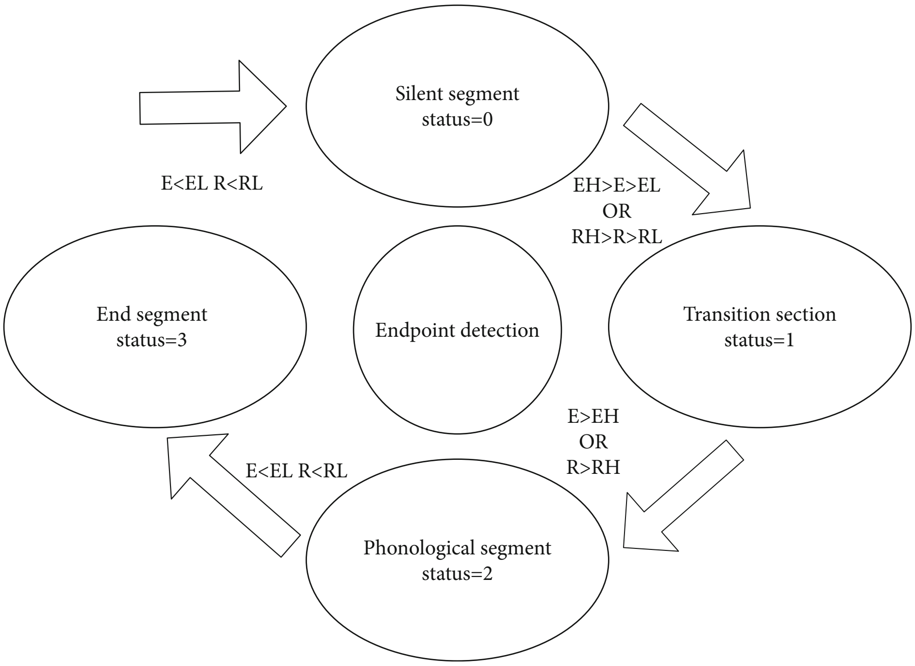

FIGURE 6: Endpoint detection state transition.

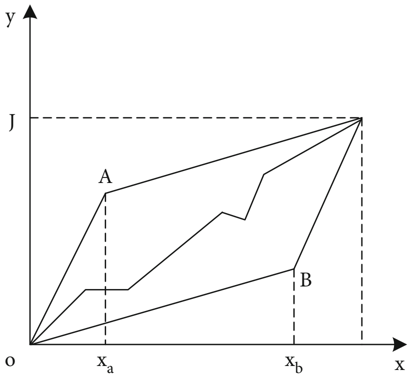

FIGURE 7: Efficient DTW algorithm.

distinction between the two categories and judge the influence of features on the recognition of stressed syllables. The higher the Fisher Ratio numerical results, the stronger the ability to distinguish stressed syllables:

$$\gamma_{\text{Fisher}} = \sigma_{\text{between}} + \sigma_{\text{within}}. \tag{22}$$

Divergent summation of such feature data:

$$\sigma_{\text{between}} = \sum_{i=1}^{c} \left(m_k^{(i)} - m_k \right)^2. \tag{23}$$

Divergent sum of such characteristic data for each class of samples:

$$\sigma_{\text{within}} = \sum_{i=1}^{c} \left[\frac{1}{n_i} \sum_{c \in \omega_i} \left(c_k^{(i)} - m_k^{(i)} \right)^2 \right]. \tag{24}$$

We tested eight-dimensional feature vectors and measured the recognition ability of stressed syllables of each feature, judge by Fisher Ratio values in training set and test set.

(1) After testing, we can find that Eavg in the third dimension and FDchr in the eighth dimension have higher Fisher Ratio, both of which are above 0.8. Secondly, the second dimension and the fourth dimension have higher values, and the Fisher Ratio values of Emax and Echr characteristics are both above 0.7. The remaining characteristic values are very low. For individual words, the ability to distinguish stressed syllables from unstressed syllables is poor, as shown in Figure 9:

(2) After the stress recognition process of the first layer, the sentence stress recognition of the second layer is carried out, and these eight feature types are still tested. We can see that this time, the highest Fisher Ratio value in the test set is the first dimension, with a value as high as 0.4 duration. The highest value in the training set is the eighth dimension, and Fisher Ratio value is as high as 0.16. Different test sets and training sets have different influences on stressed syllables. This means that the recognition ability of stressed syllables in these two dimensions is very strong. However, the recognition ability and effect of other dimensions are weak, as shown in Figure 10:

4.2. Endpoint Detection Comparison. On MATLAB platform, based on efficient DTW speech recognition system, the experimental environment is tested under three conditions: no noise, 5 dB, and 10 dB. In the experiment, 20

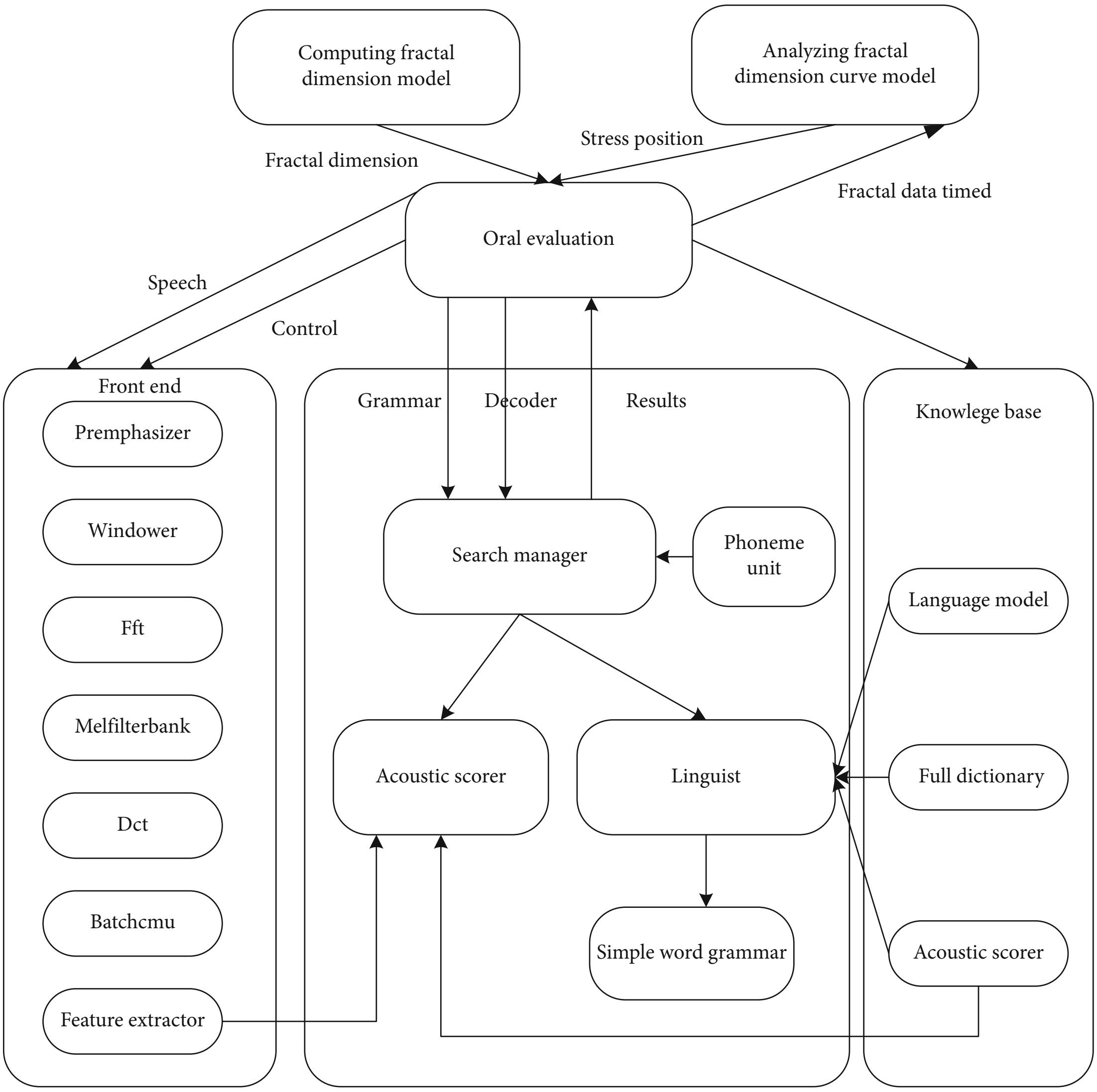

Figure 8: Improved system architecture framework based on Sphinx-4.

volunteers were invited to record the detected voice data. Among them, men and women are 50%, and each person has 10 phonetic materials. The recognition performance of dual-threshold speech endpoint detection algorithm is tested. In order to better illustrate the advantages of the detection algorithm selected in this paper, we will also use the traditional endpoint detection algorithm for data comparison. The recognition rate calculation formula is as follows:

$$\text{Recognition rate} = \frac{\text{Total number of sounds-number of misjudged sounds}}{\text{Total speech number}}. \tag{25}$$

After simulation experiment, the recognition rate of the two algorithms decreases with the increase of noise. In the same environment, the recognition rate of polysyllabic syllables is generally slightly lower than that of monosyllabic syllables. In the case of no noise, the recognition rates of the algorithms are not much different. However, with the increase of noise environment, the syllable recognition rate of this algorithm is obviously improved, and the effect is obviously superior to the traditional algorithm, as shown in Figure 11:

4.3. Stress Syllable Recognition

4.3.1. Recognition of Stressed Syllables Based on Linear Discrimination. We tested eight different phonetic features to study which features are the most effective for recognizing stressed syllables. Among them, the evaluation standard of experimental results is the recognition rate of word stress, as shown in Figure 12:

The average horizontal line is set at 70%. It can be found from the figure that the minimum and maximum normalization methods are all higher than the horizontal values, and the results are relatively balanced, so it is impossible to compare the vowel phonemes of a word. While MS normalization method is obviously superior, each feature presents a good degree of discrimination, but their recognition rate of

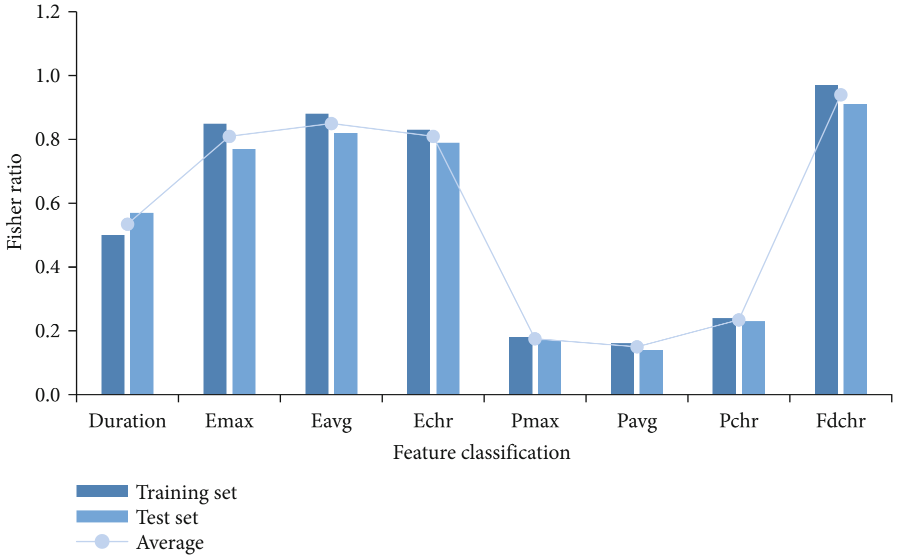

Figure 9: Word stress recognition per dimensional feature.

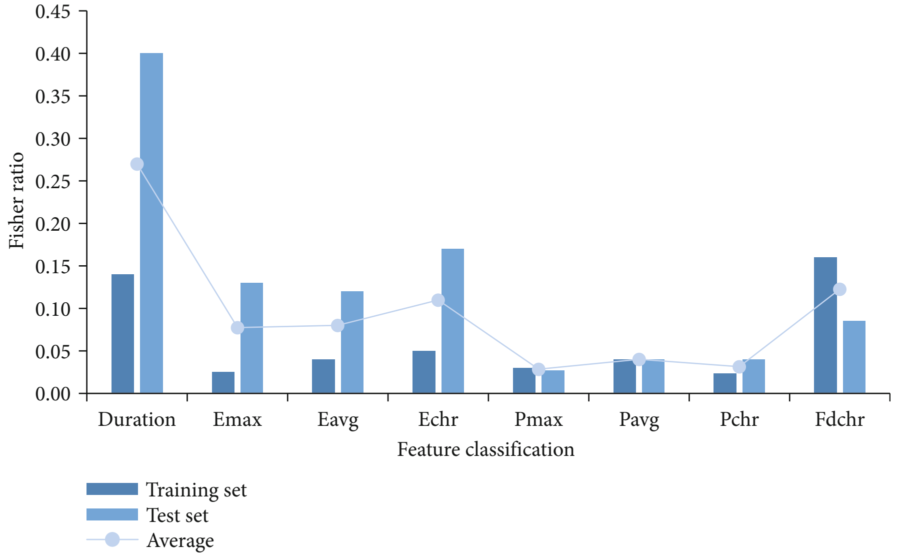

Figure 10: Sentence stress recognition per dimensional feature.

stressed syllables is very low, which cannot reflect the situation of stressed syllables well.

4.3.2. Recognition of Stressed Syllables Based on RankNet. In view of the poor results of the former method, we choose the RankNet model established in this paper, which integrates various features instead of adopting a single feature recognition method, so as to compare vowel phonemes of words and improve the recognition rate of stressed syllables, as shown in Figure 13:

The recognition rate of stressed syllables with fused features has been well improved. The error rates of the two methods are 20.6% and 19.32%, respectively. The robustness of the model is good.

Using InwMS normalization method, if any feature is removed, the error recognition rate of the model will increase by at least 3%. When dealing with duration, Eavg and FDchr, the error rate is the highest, which exceeds 24%. It can be seen that the recognition distinction of a single feature recognition is not necessarily the same in feature fusion, and their position is not necessarily important. Features are complementary, as shown in Figure 14:

4.3.3. Sentence Stress Recognition Based on RankNet. Under INwMS normalization method, the error recognition rates of accented syllables of A, B, and C are compared. A stands for R_{error}; B stands for RF_{error}; C stands for RF^2_{error}. On the test set, the reread recognition error rate of RankNet is 42.51%, as shown in Figure 15:

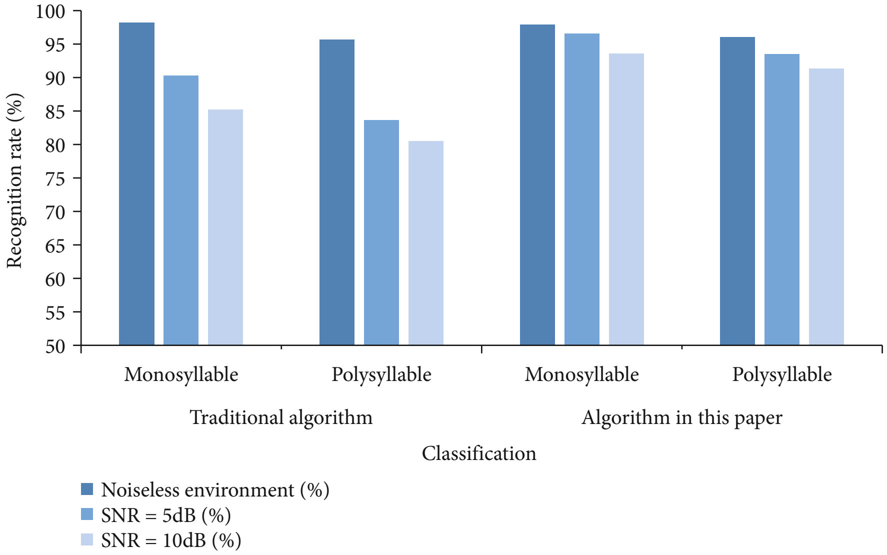

Figure 11: Recognition rate of two endpoint detection algorithms.

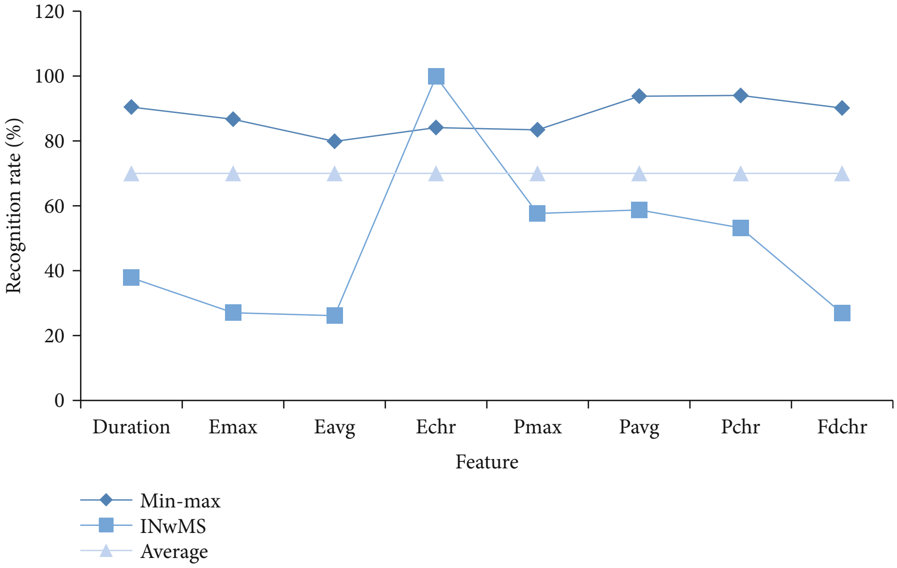

Figure 12: Recognition result of linear discrimination.

5. Conclusion

With the rapid increase of business cooperation, there are more and more people who need spoken English. However, the uneven language teaching is extremely lacking, and there is a big blank market for the system software used for oral English assistance. The research results of this paper show that

(1) in feature selection, the higher the Fisher Ratio is, the better it can distinguish stressed syllables

(2) the double-threshold speech endpoint detection algorithm is obviously superior to the traditional algorithm. Among them, the recognition rate decreases with the increase of environmental noise; the recognition rate of polysyllables is slightly lower than that of monosyllables

(3) min-max method cannot compare the vowel phonemes of words; although InwMS method can distinguish features better, the recognition rate of stressed syllables is low. Therefore, the recognition result of linear judgment of a single feature is of little significance

(4) testing the fusion features improves the recognition rate of stressed syllables. The error rates of the two normalization methods are 20.6% and 19.32%, respectively

(5) if any feature is removed, the error recognition rate of the model tested for fusion features will increase

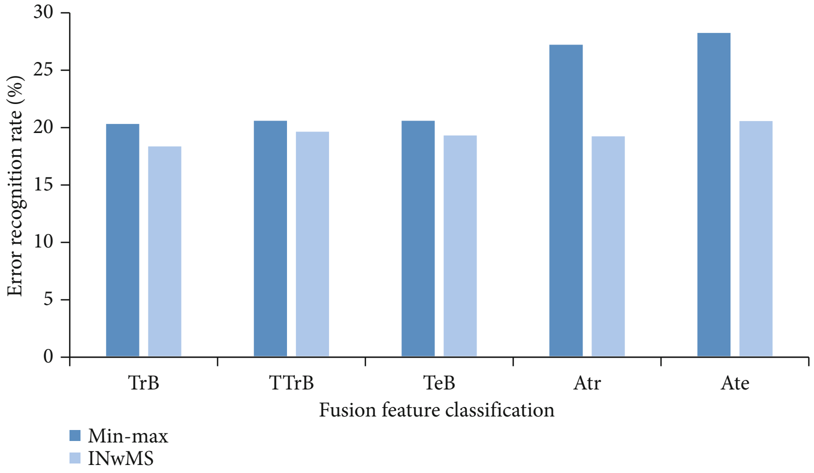

FIGURE 13: Recognition of stressed syllables by fusing features.

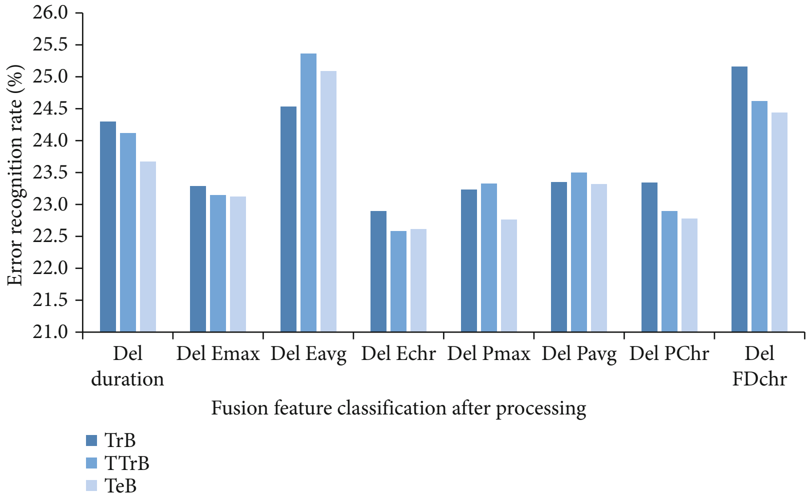

FIGURE 14: Fusion feature-recognition result of removing a feature.

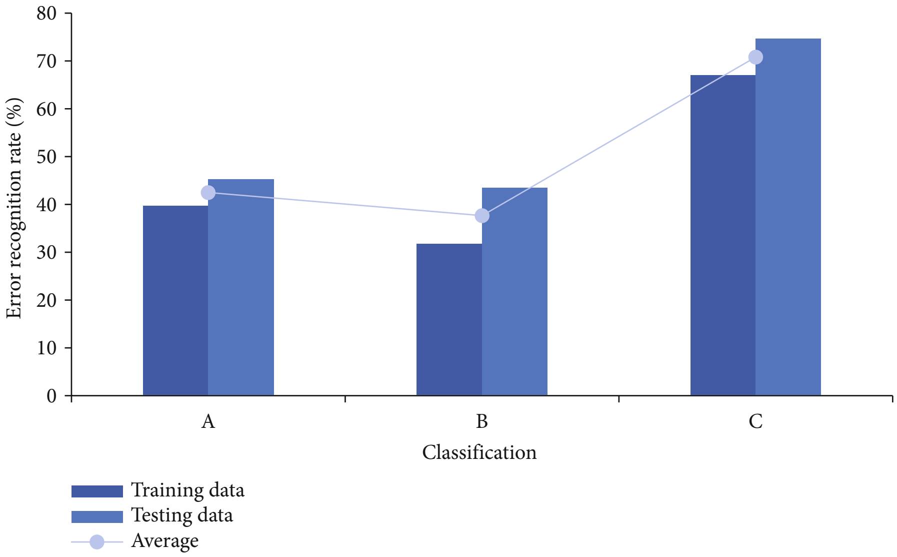

FIGURE 15: Sentence stress recognition result of fused features.

by at least 3%. After individual features are removed, the error rate of feature recognition rises linearly

(6) under INwMS normalization method, for sentence stress recognition based on feature fusion, the rereading recognition error rate of RankNet on the test set is as high as 42.51%

The established RankNet algorithm model has stable recognition effect and good reliability. The final experimental training is a good practice of combining experiment with theory, which has more obvious advantages than other traditional methods. The method in this paper still has some shortcomings. For example, the phonetic corpus is small in scale; there will still be missed judgments and misjudgments in noise environment. In the future work, we try some ways to further improve the recognition performance, for example, develop distributed systems to reduce maintenance costs; increase the size of the corpus to make the phonetic representation more extensive; further expand and refine the grammar, rhythm, and rhythm of stressed syllables; improve the speech endpoint detection algorithm to overcome the influence of noise. With the continuous research of speech-related topics, its development will be constantly innovated with the development of the times, and more new ideas will be put into practice to strive for breakthrough progress.

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 authors declare that they have no conflicts of interest regarding this work.

References

[1] S. Lee, Y. Lee, and Y. Kwon, "The effect of syllable and morpheme on visual word recognition: comparison between simple words and compound words," *Journal of Language Sciences*, vol. 27, no. 4, pp. 75–96, 2020.

[2] A. Winkler, R. Carroll, and I. Holube, "Impact of lexical parameters and audibility on the recognition of the Freiburg monosyllabic speech test," *Ear and Hearing*, vol. 41, no. 1, pp. 136–142, 2020.

[3] C. Moore-Cantwell, "Weight and final vowels in the English stress system," *Phonology*, vol. 37, no. 4, pp. 657–695, 2020.

[4] C. Kaland, "Offline and online processing of acoustic cues to word stress in Papuan Malay," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. 731–747, 2020.

[5] S. Bakst and C. A. Niziolek, "Effects of syllable stress in adaptation to altered auditory feedback in vowels," *The Journal of the Acoustical Society of America*, vol. 149, no. 1, pp. 708–719, 2021.

[6] V. Y. Yu, "Effects of syllable position, fundamental frequency, duration and amplitude on word stress in Mandarin Chinese," *Journal of Psycholinguistic Research*, vol. 50, no. 32, pp. 293–312, 2020.

[7] B. Lrincz, "Concurrent phonetic transcription, lexical stress assignment and syllabification with deep neural networks," *Procedia Computer Science*, vol. 176, pp. 108–117, 2020.

[8] Q. Zhang, "Recognition of English spoken stressed syllables based on natural language processing and endpoint detection algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 4, pp. 5713–5724, 2020.

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

[10] L. Cilibrasi and V. St Ojanovik, "The interplay of stress saliency and word beginning saliency: an experimental study," *Linguistica Pragensia*, vol. 30, no. 2, pp. 113–126, 2020.

[11] M. S. Ahmed, "The impact of students' realization of stress placement on identifying the class of the English word and its meaning," *International Journal of Psychosocial Rehabilitation*, vol. 24, no. 10, pp. 3707–3719, 2020.

[12] X. Zhu, "Hybrid algorithm of English speech recognition based on hidden Markov model and clustering," *Computer Measurement and Control*, vol. 28, no. 5, p. 5, 2020.

[13] C. Xiaohong and T. Hua, "Research on English speech recognition based on deep machine learning," *Journal of Guiyang University: Natural Science Edition*, vol. 16, no. 3, p. 5, 2021.

[14] L. Xia and T. Gaofeng, "Research on English acoustic detection system based on speech recognition," *Automation Technology and Application*, vol. 38, no. 12, p. 3, 2019.

[15] W. Qionghua, "An empirical study on the application of speech recognition in college English listening teaching," *Research on Communication Power*, vol. 4, no. 24, p. 2, 2020.

[16] A. L. Agostinho and L. M. Hyman, "Word prosody in Lung'Ie: one system or two?," *Probus*, vol. 33, no. 1, pp. 57–93, 2021.

[17] A. Mai, "Phonetic effects of onset complexity on the English syllable," *Laboratory Phonology*, vol. 11, no. 1, p. 4, 2020.

[18] A. K. Emmendorfer, J. M. Correia, B. M. Jansma, S. A. Kotz, and M. Bonte, "ERP mismatch response to phonological and temporal regularities in speech," *Scientific Reports*, vol. 10, no. 1, p. 12, 2020.

[19] A. Kukhto, "The acoustics of word stress in Gaeilge Chorca Dhuibhne," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2726–2726, 2020.

[20] E. Banzina, L. C. Dilley, and L. E. Hewitt, "The role of secondary-stressed and unstressed–unreduced syllables in word recognition: acoustic and perceptual studies with Russian learners of English," *Journal of Psycholinguistic Research*, vol. 45, no. 4, pp. 813–831, 2016.

[21] L. Lei, "Design and research of embedded English speech recognition control system for intelligent seeder," *Agricultural Mechanization Research*, vol. 40, no. 12, p. 5, 2018.

[22] L. Lin, "Research on the application of automatic speech recognition technology in English phoneme teaching for non-English majors," *Journal of Kaifeng Institute of Education*, vol. 38, no. 4, p. 2, 2018.

[23] H. Yunxiao, S. Qing, F. Yuxiang, and G. Qing, "Speech endpoint detection algorithm based on MFCC distance in complex noise," *Computer Engineering*, vol. 46, no. 3, p. 6, 2020.

[24] D. Honghu, Q. Mao, and Z. Xiaobing, “Improved endpoint detection algorithm based on LMS adaptation,” *Journal of Changzhou Institute of Technology*, vol. 34, no. 1, p. 10, 2021.

[25] J. Yu and Z. Xiaoqun, “Research on endpoint detection algorithm applied to speech annotation in noisy environment,” *Journal of Nanjing University of Posts and Telecommunications: Natural Science Edition*, vol. 41, no. 1, p. 9, 2021.