Hindawi
Complexity
Volume 2023, Article ID 9815140, 1 page
https://doi.org/10.1155/2023/9815140

Retraction

Retracted: English Audio Language Retrieval Based on Adaptive Speech-Adjusting Algorithm

Complexity

Received 19 December 2023; Accepted 19 December 2023; Published 20 December 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] X. Feng and Y. Zhou, "English Audio Language Retrieval Based on Adaptive Speech-Adjusting Algorithm," *Complexity*, vol. 2021, Article ID 2762180, 12 pages, 2021.

Hindawi
Complexity
Volume 2021, Article ID 2762180, 12 pages
https://doi.org/10.1155/2021/2762180

Research Article

English Audio Language Retrieval Based on Adaptive Speech-Adjusting Algorithm

Xiaoyan Feng [1] **and Yanfang Zhou** [2]

[1]*School of Foreign Languages, Ningxia Normal University, Guyuan, Ningxia 756000, China*
[2]*Department of Preschool Education, Lishui University, Lishui, Zhejiang 323000, China*

Correspondence should be addressed to Yanfang Zhou; zhouyanfang@lsu.edu.cn

Received 15 April 2021; Accepted 19 June 2021; Published 3 July 2021

Academic Editor: Zhihan Lv

For the purpose of language retrieval for English listening, this paper designs and implements a cross-language information retrieval system for English listening. Different implementation methods of cross-language information retrieval, query, and translation are analyzed. The system adopts cross-language information retrieval technology based on bilingual dictionaries. According to the cross-language retrieval system of the existing bilingual dictionaries and monolingual dictionaries, based on the design and implementation of the fuzzy search dictionary lookup mechanism, the existing dictionary lookup mechanism is constructed and analyzed. Aiming at the problem of translation ambiguity in information retrieval systems based on bilingual dictionaries, a disambiguities elimination algorithm based on cooccurrence technology is proposed. In continuous speech, the speed of different speakers in different contexts is very different. Deviation from normal speech speed often leads to recognition errors, which makes recognition performance decline. Considering that the influence of speech speed on the length of speech units increases or decreases synchronously, and there is a strong correlation between the lengths of adjacent speech units, an adaptive speech speed algorithm is proposed based on the framework of implicit Markov model based on the information of the length of speech units. Experiments on number string and large vocabulary continuous speech recognition show that the algorithm is effective.

1. Introduction

In continuous speech, the speed of speech of different speakers in different contexts is very different. Excessive deviations from normal speech speed tend to cause recognition errors [1]. Excessive speech speed will increase deletion errors, while too slow speech speed will increase insertion errors, thus degrading recognition performance. Such a problem has attracted the attention of researchers, and some achievements have been made on how to reduce recognition errors and improve system performance under the condition of deviating from normal speech speed [2]. Overall research on this problem at present is mainly according to some way to get first identify the corpus of speed measurement, and then according to the speed of fast update transition probability, and in the case that the state becomes slow, because the transition probability increases, the left transition probability becomes smaller and fast; on the contrary, the duration of each speech unit can be adjusted to adapt to the speed. We note that the above research is still based on the classical HMM framework, the modified state transition probability is a time-independent quantity, and the exponential segment length distribution implied by the homogeneous Markov process does not change. Therefore, in theory, the performance improvement that can be obtained by adjusting the speed of speech by modifying the transition probability is very limited.

In the process of the development of information retrieval, people put forward various retrieval models, which have been widely applied at present; there are Boolean model and vector space model [3]. There are many variations according to these models, such as extended Boolean model [4], generalized vector model [5], latent semantic indexing model, and neural network model. The most

widely used in it is the vector space model; the vector space model has many advantages. Index term weighting improves the retrieval effect, and documents are retrieved according to the sorting related to document retrieval. The vector space model is a resilient sorting strategy; its performance is fairly good. Vector model is simple and convenient and has become a popular retrieval model. The cross-language information retrieval query language used by the user is called the source language. Taking the retrieval document language as the target language, building a communication bridge is the core of cross-language information retrieval technology and a key issue in research. Because the cross-language information retrieval in the processing of the object is different, cross language information retrieval technique can be divided into two types: document translation and query translation [6]. Speech length adjustment refers to the changing voice playback speed while maintaining the perceptual characteristics of voice, such as pitch, and formant structure. This makes the adjusted voice sound as if the speaker himself is changing the speed of his speech. Adjustment has been widely used in language teaching, speech coding, speech synthesis, and audio retrieval [7]. In the literature [8], a simple length adjustment algorithm overlap-adding method (OLA) is put forward. It uses window function one by one to divide the original voice signal into a series of overlapping frames, by changing the length of the overlap to realize the voice signal compression or extension, but this algorithm does not take into account the continuity of adjacent frames. It is easy to cause pitch distortion and back holes. In order to solve this problem, the wave superposition method [9] is proposed. When determining the stacking frame position, a long deviation is introduced to superimpose consecutive frames together to ensure the continuity of the frame signal superimposition. However, WSOLA algorithm adopts fixed adjustment factor. It is easy to lose your voice turning part in compression and phoneme and intonation when extended to fuzzy, which are not easy to identify. Voice signal is complex and changeable; one of the important factors is the speed of change. Although the hidden Markov model in the state has a certain degree of freedom in time [10] when the speed change is dramatic, system identification performance is significantly lower [11]. Aiming at the problem of poor adaptability of the speaker's variable speed model, the usual approach is to adjust the speed through the speaker and statement level. By changing the phase of long frame/frame interval feature extraction, the adjusted phonemes of consecutive frames are kept within a certain range. Three groups were selected, and the frame length and frame interval were moved to adjust the speed, which increased by 2% in the recognition result [12]. Based on the word success, discriminative training is applied to the speed detection and then adjusted to three setting rates, which improves the phoneme recognition rate by 1.7% [13, 14]. SPE is described emphatically from the angle of the vocal voice production process. GP is through the acoustic spectrum analysis and draw out the phonemes in 11 kinds of complementary components, through the combination of these components can be described and distinguish all the phonemes [1, 15]. To a certain extent, the redundant information and noise suppression characteristics [16] to identify units related to distinguish sexual information is retained in the probabilistic event [17, 18]. It is not hard to find the above several adjustment methods using only several fixed, long discrete values of the frame and frame interval of speed adjustment; adaptation to the characteristics of the speech signal rate varied [6]; in addition, by means of identification results, or other system model training methods to measure the speed of the detection speed, increase the burden of the system [19–25]. In view of this, this paper proposes an adaptive length adjustment algorithm. First, the voice signal is divided into a series of stable speech segments. Each voice and a phoneme or basically corresponding syllables are produced. Then according to the total adjustment factor and the segment speech length distribution of each speech segment, partial adjustment is made, and finally the speech length is adjusted by the WSOLA algorithm.

This paper first introduces the classical information retrieval model and then analyzes the model to be adopted for the text retrieval at the sentence level, designs the data structure of the index and the way of English audio storage, and finally implements the English audio retrieval system at the sentence level for mobile platform. Hidden Markov model based on extended directly from long, with a state resident length probability instead of state transition probability as the basic parameters, and the changes in the velocity of the speech directly reflect the change of for long, at the same time in the long effects of synchronous growth speed changes or synchronous decline, namely under the condition of low speed, a phonetic unit is longer than the average before long, after a speech unit will be in the same trend is longer than the average long, and fast under the condition of the opposite. And in a short period of time, a speaker's speaking speed will be relatively stable; that is, the influence of such speed on the length of a paragraph can be considered to be basically the same. In this way, the deviation of the previous segment length from its mean value can be used to predict the changing trend of the later segment length.

2. Research on Adaptive Speech-Adjusting Algorithm for English Listening and Audio

2.1. Adaptive Speed Adjustment for English Audio Language Retrieval. The method of time-domain analysis is directly related to the time-domain waveform of speech signal. In the digital processing of speech signal, the first contact and the most intuitive is its time-domain waveform. But how do you make a speech signal into a time-domain waveform? The digital representation of speech signals is a fundamental problem. In this case, we rely on the well-known sampling theorem, which tells us that a band-limited signal can be represented by periodic sampling points in the time domain, provided that the sampling rate is high enough. The whole process of speech speed adjustment is as follows: first, the speaker is used to input the speech data to be recognized;

after reading the data in the data area, the short-time average amplitude of the speech signal is calculated; then, the vowel starting and ending points of the speech data are analysed according to the short-time average amplitude; finally, the speed adjustment is decided according to the length of the vowel. If the vowel length is not within the specified recognition range, it is adjusted; otherwise, it is not adjusted. The overall block diagram of speech speed adjustment is shown in Figure 1.

Due to the quasistationary property of speech signal, any digital processing algorithm and technology of speech signal are based on "short time." In order to achieve various speech adjustments, some examples of speech signals described in time-domain metrics, including short-term energy, short-term mean amplitude, and short-term zero-crossing rate, must be understood. These descriptive methods are good because the required digital processing is very simple to implement and provide a useful basis for estimating important characteristics of speech signals. These short-term parameters of speech signal are described in detail in the following sections.

Considering the short-time stationary characteristic of speech signal, speech signal should be processed in sections. The segmentation of signal stream is realized by using the method of weighting of movable finite length window; that is, a segment of speech $S(n)$ is windowed, and a window function $W(n)$ is multiplied by $S(n)$ to form windowed speech. The window functions commonly used in digital processing of speech signals are square window and Hamming window, and their expressions are as follows (where N is the window length):

The square window:

$$\delta(n) = \begin{cases} 1, & n \in (0, N-1), \\ 0. \end{cases} \tag{1}$$

Hamming window:

$$\delta(n) = \begin{cases} 0.5 + 0.6\cos\dfrac{n}{N-1}, & n \in (0, N-1), \\ 0. \end{cases} \tag{2}$$

In the calculation of the three short-term parameters mentioned above, the square window or Hamming window is generally used. These short-time processing techniques can be expressed in mathematical form:

$$Q = \sum_{k} T(x(k))\delta(n-k). \tag{3}$$

The speech signal (or the required frequency band filtered by linear filtering) is transformed by T, which can be linear or nonlinear, depending on a tunable parameter or a set of parameters. The resulting sequence is then multiplied by the window sequence at the time consistent with the sampling flag, and the sum of all nonzero values of the product is finally concluded. Usually the width of the window sequence is finite, so it is the sequence of partial weighted average values of the sequence.

In the standard Boolean model, audio is expressed as follows:

$$V_i = \{T_{i1}, T_{i2}, \ldots, T_{in}\}, \tag{4}$$

$$\mathrm{sim}(t_i, q) = \frac{P(R \mid t_i)}{P(R' \mid t_i)}. \tag{5}$$

The similarity between the sentence and the query is calculated by the inner product operation of two vectors. The inner product calculation method of the similarity between the query and the sentence is

$$\mathrm{sim}(t_j, q) = \sum_{i} t_{iq} t_{ij}, \tag{6}$$

$$\mathrm{sim}(t_j, q) = \frac{2\sum t_{jq} t_{ij}}{\sum t_{jq}^2 \sum t_{ij}^2}. \tag{7}$$

2.2. An Adaptive Speech Speed Algorithm Based on Segment Length Correlation. If the speech speed deviating from the training data is too fast or too slow, there will be a large mismatch between the corresponding segment length and the segment length information trained by the training corpus, which will degrade the recognition performance. In fact, in continuous speech, the change of speed can be reflected as the change of length. In the case of fast speech speed, the length of the corresponding speech unit segment will be longer than the average segment length, while in the case of slow speech speed, the opposite is true. Moreover, we believe that the segment length of adjacent speech units is greatly correlated with the deviation degree of the mean value of each segment length due to the influence of speech speed. Therefore, if the unary probability system of segment length is reasonably modified and relevant information of segment length is included, it should be an effective way to solve the problem of speech speed.

The speech length adjustment is divided into time-domain compression and time-domain extension. Time-domain compression is mainly used for audio retrieval and voice mail and other purposes, hoping to maintain a high quality of speech in the case of large compression ratio. Time-domain expansion is mainly used in language teaching, especially in foreign language learning, where the faster the speed is required to play the slower.

Based on the above practical applications, the speech signal is divided into a series of stable speech segments by the algorithm in this paper. When each segment is compressed relative to a phoneme or syllable in the time domain, the longer the speech segment is, the more stable the characteristics are and the more similar the waveform is, so it can be compressed to a greater extent without losing too much information. Shorter speech segment is often the turning part of speech, with a short duration and a smaller proportion of compression, which can avoid the loss of information. In the time-domain expansion, short speech segments are shown to speak faster in auditory sense, and the speech speed of longer speech segments is relatively slow

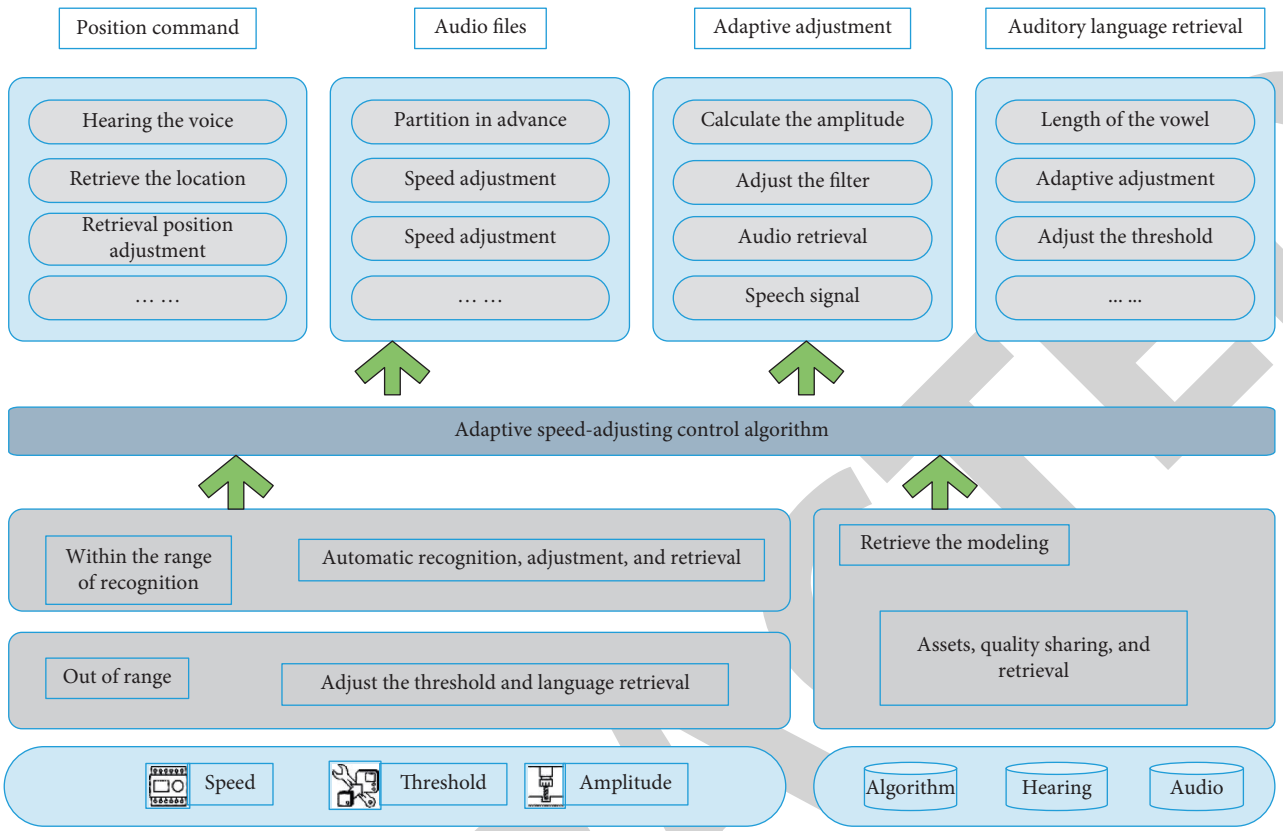

Figure 1: Block diagram of self-adaptive adjustment of speech speed.

when extended to a larger extent, and the speech speed of longer speech segments can be expanded to a smaller extent to avoid the ambiguity of phonemes and intonation. The distribution flowchart of adaptive local adjustment factors is shown in Figure 2.

In order to improve the execution efficiency of the algorithm, the speech segment of the local adjustment factor will not be divided into frames, but the whole speech segment will be directly superimposed. In order to ensure the continuity between synthesized speech segments, the first frame of each speech segment is selected according to the last frame of the previous speech segment, and the selection range is in the neighborhood of the starting point of the speech segment.

For statement level text retrieval, all documents are at the statement level, so the vector dimensions represented by sentences are not very large. According to the statistics of 550 sentences in the corpus (the statistical information is shown in Table 1), it is found that the longest sentence only contains 31 valid words, and the average number of valid words in each sentence of 550 sentences is 34.

Therefore, the similarity between the vector and the query can be calculated directly by using the formula, without considering the dimension compression of the vector. The vector space model thinks that lexical items are independent without considering the semantic and positional relations between lexical items, which is the deficiency of vector space model in the field of information retrieval. But in English audio retrieval systems, the independence and neutrality position happens to be a system, the advantages of the system are English listening retrieval, the user's retrieval words need a lot of time in the pronunciation of the word, rather than the semantic relationships of words, so in this case using the vector space model to statement of English audio retrieval system is completely feasible.

Short-term mean energy is defined as

$$E_I = \sum \delta^2(k) t(n-k). \tag{8}$$

When the window starting point is $n = 0$, the short-time average amplitude of the speech signal is calculated as

$$M_0 = V(1) + V(2) +, \ldots, V(n). \tag{9}$$

For broadband signal, in order to reflect the variation of its zero-crossing rate with time, the long-term average zero-crossing rate cannot be adopted, but the short-term average zero-crossing rate must be adopted, which is defined as follows:

$$V_N = \sum \operatorname{sgn}\delta(k) t(n-k), \tag{10}$$

$$T_N = \operatorname{sgn}\delta(k) - \operatorname{sgn}\delta(k-1). \tag{11}$$

In the field of information retrieval, inverted file is the most widely used and most effective way of the index system using inverted file to create English text index database, at the same time the corresponding English audio information path shall be stored in inverted file, so that when the user

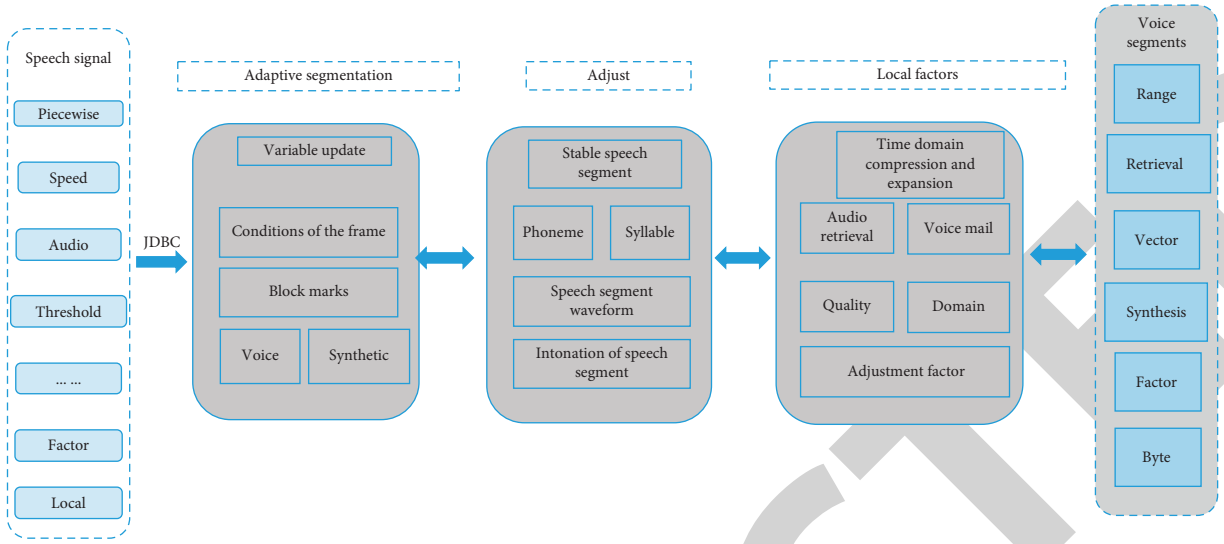

Figure 2: Distribution flowchart of adaptive local adjustment factors.

Table 1: Sentence set statistics.

The total number of sentences	The shortest sentence		The longest sentence		Average	
	Number of bytes	Number of words	Number of bytes	Number of words	Average length	Average number of words
550	35	10	195	35	32	18

input keywords retrieval English audio information, the realization of the first is the knowledge of the English text retrieval, and then through the text information retrieval of positioning to the corresponding audio information back to the user. Since there are only 26 letters in English, our index organization structure is distributed into 26 files according to the method of hash. The data structure of the index contains the following information, as shown in Algorithm 1.

Every time the indexing program reads a file that needs to be indexed, it first divides the contents of the file into words, restores the word form to the result of the word segmentation, and then calculates the position of the word in the index with the hash function. If no data exists at the location, write it. If there is already data in this location, then compare the two words and whether the file name is the same. If so, change the occurrence number of the word in the index entry and write it back to the index. If it is different, follow the address of the next item in the current position to find the end of the chain, write the index item to the end of the file, and record its address to the last item in the current chain.

3. English Listening Audio Retrieval Based on Adaptive Speech Speed Adjustment

As the speech signal has short-term stationary characteristics, the first-order and second-order statistical characteristics remain unchanged in the short time window, so the frame length and frame shift interval are usually selected as 5 ms, respectively, in the experiment, which is based on the compromise between its statistical characteristics and empirical values. However, the use of fixed frame length and frame shift interval is not the optimal choice; some of the phonemes in the stop duration may be less than 5 ms, and some vowels may be more than 100 ms, and different sentences in terms of speed are also very different. In this paper, the average duration of phonemes is used to measure speech speed. Here, sentences are taken as the unit. Figure 3 shows the statistic histogram of speech speed of statements in TIMIT corpus. The abscissa axis is the duration of phonemes, and the ordinate is the number of corresponding statements.

It is not difficult to find from Figure 3 that the speed of different sentences in the corpus varies greatly. The sentences with an average phoneme duration of 6 ms account for more than 90% of the whole corpus, and the sentences with an average phoneme duration of 4.2 ms are the most. In order to reduce the negative impact of speech speed change on recognition performance, existing algorithms usually adjust speech speed at the speaker, sentence, or phoneme level. By adjusting the frame length and frame shift interval in the preprocessing stage, the average number of lasting frames of phoneme is relatively constant. According to the existing experiments, it has been proved that the adaptive adjustment algorithm based on statements has the best effect. Therefore, the method in this paper also chooses statements as the basic adjustment unit.

How to accurately and reliably detect the speed of speech is the key problem for speed adjustment. It is worth noting that mute and other non-speech segments are removed here. Assuming that a given statement i consists of n words, the average duration of each phoneme in the statement is

```
Struct indexNode
{
Word/words
Weight//word weight
Filename//text filename, a common prefix for text and speech
Soundfile//audio file number
Soundindex//offset in audio file
SoundLength//the length of the data in the audio file
Textfile//textfile number
TextIndex//offset in the text file
TextLength//the length of the data in the text file
Next//next address
}
```

Algorithm 1: The structure of audio language inodes.

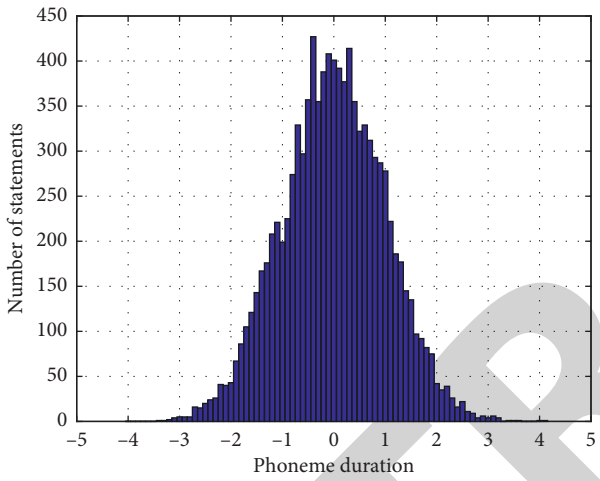

Figure 3: Histogram of speed statistics.

$$T(i) = \frac{\sum_j t(\delta_i)}{\sum_j n(\delta_{i,j})}. \tag{12}$$

The goal of adjustment is to make the speed of each sentence consistent with the average rate of the corpus; that is, the average duration of phonemes is consistent. T is defined as the average phoneme duration of the corpus; then T can be expressed as

$$\phi = \frac{\sum_i \sum_j t(\delta_{i,j})}{\sum_i \sum_j n(\delta_{i,j})}, \tag{13}$$

where M is the total number of statements in the training set.

TIMIT corpus contains annotated files of words and phonemes. In the training stage, the number of phonemes in each statement and the duration of the statement can be obtained through these annotated files, so as to calculate the rate $f(I)$ of each statement. However, during the test phase, no information can be retrieved from the test set annotation file. Therefore, the number and time information of phonemes are obtained from the recognition results of the existing system, so as to obtain the speed of the test set, which undoubtedly increases the workload and requires the help of other systems. Since the experiment in this paper includes two stages: calculation and modeling of a posteriori probability of phoneme attributes, and phoneme attributes and phonemes correspond one-to-one, the boundary of phonemes can be obtained from the posterior probability information, so as to obtain the number of phonemes.

It can be found from the experiment that the angle value between different classes of vectors is large, about 70–90 degrees. However, the angle value between vectors of the same category is relatively small, generally less than 30 degrees. In this paper, the following three steps are adopted to detect the boundary.

(1) The posterior probability values of phoneme attributes in the mute frame are all small. In addition, the mute segment cannot contain phoneme boundaries, and the detection rate is high. Therefore, the mute segment is first determined and removed. The specific method is as follows. If the posterior probability of a certain frame's phonemic attribute is less than 0.1 and lasts for more than three frames, this part of the frame is determined as a silent segment.

(2) Taking 30 degrees as the threshold, the maximum point with an angle greater than the threshold value was selected as the candidate boundary. Although there were many false boundary points in the selected points, they could be effectively removed by decision rules in the following steps.

(3) Among the candidate boundaries selected in Step (2), only the frames with sharp peak values are likely to be boundaries. Therefore, this paper proposes the following two necessary conditions for boundary screening: (1) the included angle value of the extreme point is more than 2% larger than the included angle value of the preceding and the following two frames; (2) the angle value of the extreme point is more than 10% greater than the minimum value within 5 frames before and after. These two conditions can determine the sharpness near the extreme point.

When calculating the boundary detection score, as long as the deviation in the boundary and corpus annotation files is no more than 3 frames, the detection is considered correct. The detection results are shown in Table 2.

Table 2: Boundary detection results based on a posteriori probability vector of phonemic attributes.

	All the border	To detect the boundary	Remove the border	Into the boundary
Number	48995	41692	7303	8526
Percentage	100%	87.3	12.8%	19.5%

As can be seen from Table 2, the detection rate of the boundary is 87.3. Since this experiment only needs to get the number of phoneme boundaries and does not care about the exact boundary position, the number of insertion errors can offset part of the influence of deletion errors. Therefore, the detection results can meet the needs of speech speed estimation.

4. Experimental Verification

The algorithm in this paper conducts experiments on TIMIT corpus. 35% statements from the recommended TRAIN set in the corpus are selected as the training set, and 1348 statements from TEST are selected as the TEST set, and there is no overlap between the TRAIN set and the TEST set. In addition, the statements in SA1 and SA2 are excluded. TIMIT contains 65 phoneme units, and according to the standard of IPA division, there are 47 phonemes when the phonemes are converted into phonemic properties. The phoneme list is mapped to the CMU/MIT-39 phoneme set to evaluate the recognition performance.

Since phonemic attributes can be regarded as distinguishing features of phonemes, they are usually used as feature vectors for back-end modeling. In the detection of phonemic attributes, the output is the posterior probability of the corresponding phonemic attributes. In order to make it more discrete, this paper carries out logarithmic operation on it. Finally, HMM model is used to model the transformed vector, and the modeling unit is the phoneme.

In this paper, a three-layer neural network is used in the detection of SPE and GP, including input layer, hidden layer, and output layer, and the hidden layer contains 350 neurons. The speech speed adjustment algorithm in this paper acts on the front-end preprocessing stage, and its acoustic features are 15-dimensional MFCC, I-dimensional energy features, and their first-order and second-order differences, a total of 39 dimensions. In order to consider the influence of long-term characteristics, the input end of the neural network in this paper is the characteristic parameters of 9 consecutive frames. Therefore, the input terminal contains $39 \times 10 = 390$ nodes, and the number of nodes at the output terminal is consistent with the number of SPE and GP. For the detection of SPE or GP attributes, a neural network with multiple output ports needs to be trained.

The speech characteristics of binary features (0 or 1) and the posterior probability of phoneme attributes corresponding to the output of the neural network (real numbers between 0 and 1) are extracted to analyze its detection performance. Here we take 0.5 as the threshold set test result of the discretization test, which is derived from the corpus of phoneme annotation file information. The results as shown in Table 3.

In Figure 4, the PESQ value corresponding to 600 BPS is the speech quality of the vocoder itself. It can be seen from the figure that the coding scheme based on the adaptive speech speed adjustment algorithm is generally better than the coding scheme based on the STFT algorithm. When the encoding rate is reduced to 300 BPS, the encoding scheme based on the adaptive speech speed adjustment algorithm still achieves a relatively satisfactory effect on the pure speech, and the synthesized speech maintains a good naturalness and intelligability.

We carried out a simulation experiment on English. English pronunciation is derived from TIMIT language library, and the sampling rate is 8 K·Hz. Figure 5 shows the performance simulation results of the algorithm in this paper when the adjustment factor $U = 2$ is used for compression and $U = 0.5$ is used for expansion of speech.

Figures 5(a) to 5(b), respectively, describe the situation of speech compression and extension: the local adjustment factor U adaptively adjusts with the speech signal, and its size is proportional to the length of the speech segment and is strictly restricted within the threshold range. The waveform of the adjusted speech is very similar to the original speech, but the proportion of each speech segment is changed. For example, due to the stability of the waveform, it is divided into a longer speech segment, and a larger adjustment factor U is obtained. Although the proportion of the waveform is reduced after compression, it still maintains a high intelligability. Due to the change of waveform, it is divided into several shorter speech segments, and smaller adjustment factor U is obtained. After expansion, each phoneme can be heard more clearly.

Can multimodal mode be used more than "English listening and audio retrieval satisfaction" in statistics? Data consolidation and analysis are shown in Figure 6.

As can be seen from the bar chart above, 42.59% of students are very fond of teachers using the multimode listening teaching mode. 49.07% of the students think that listening class teaching using PPT will be more effective. In addition, 50.93% of the students expect that teachers can use multimode teaching in English listening class.

In contrast to decreasing the pitch period, when the pitch period is smaller than P min, the pitch of the input speech is too high. The pitch of the speech should be lowered by increasing the sampling points of the speech signal (but keeping the basic waveform unchanged). Similarly, to keep the basic waveform unchanged and the speech pronunciation basically undistorted, the most important thing is to add points on the same curve between the two sampling points (interpolation). In this paper, the linear interpolation method was adopted to increase the point between two sampling points (as can be seen from Figure 7, a column of the original digital signal, a linear interpolation is calculating the midpoint between two points in the original sequence,

Table 3: Test results of SPE attributes in test set.

Phonological properties	Acc (%)
Anterior	90
Back	87.4
Consonantal	88
Continuant	92.3
Corona	89.6
High	88.1
Low	92.6
Nasal	97.5
Round	93.6
Silence	97.5
Strident	96.4
Tense	90.3
Vocalic	87.5
Voice	92.5

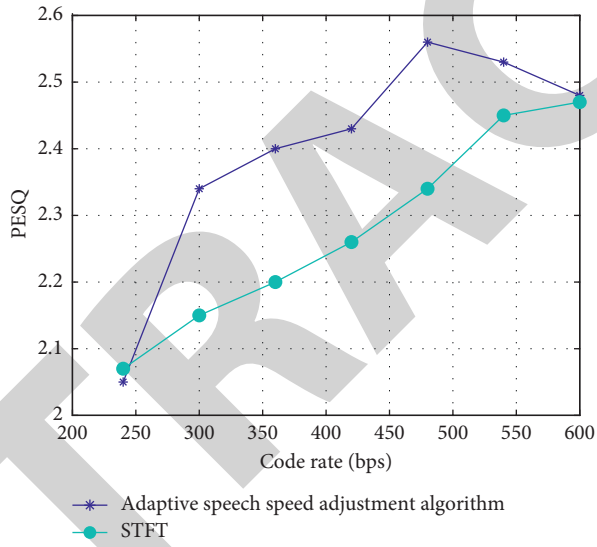

Figure 4: PESQ score corresponding to 200 BPS–600 BPS rate.

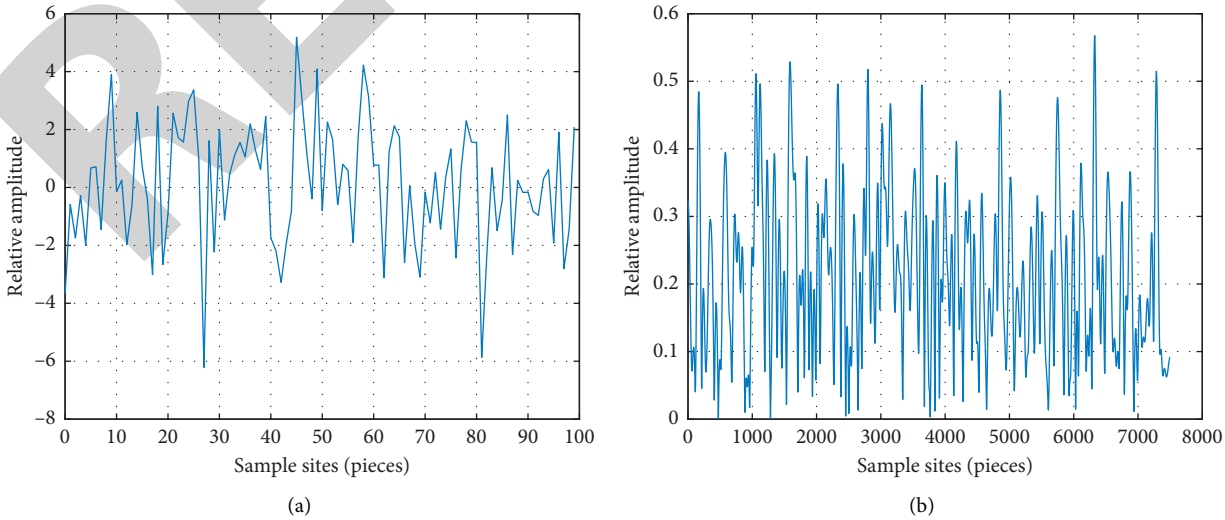

Figure 5: Speech waveform and local adjustment factors. (a) $U = 2$ compressed English listening audio retrieval performance. (b) $U = 0.5$ compressed English listening audio retrieval performance.

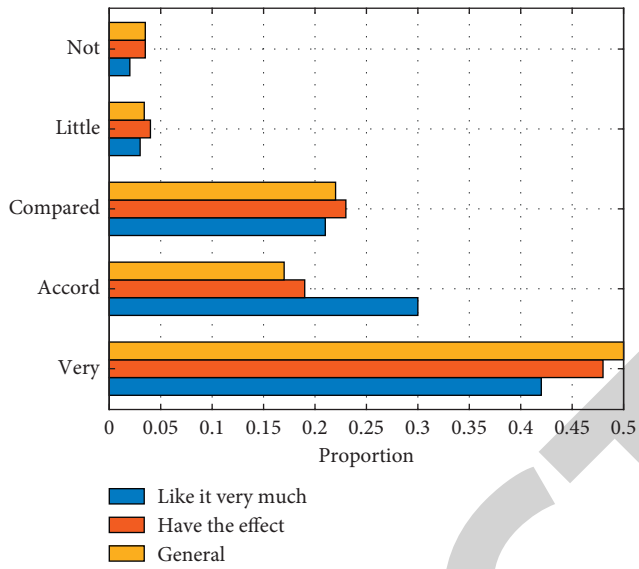

Figure 6: Feedback survey results of English listening audio language retrieval.

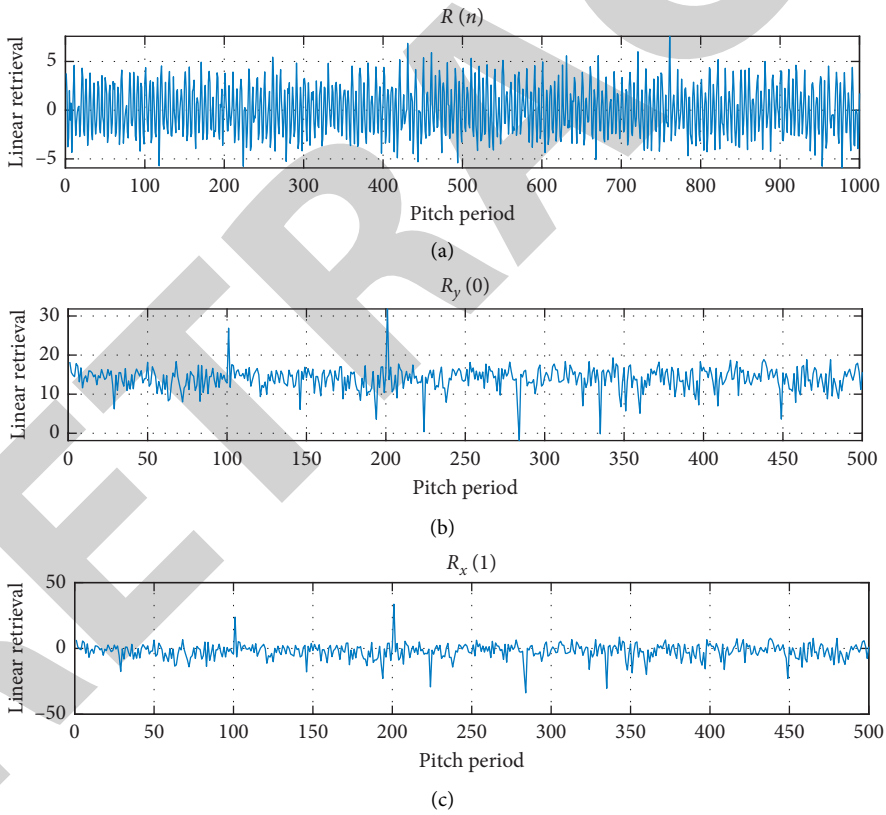

Figure 7: Retrieval effect diagram of speech signal after central clipping.

and then to get the new points in sequence, which is a one-dimensional numerical sequence amplification in spatial domain, or there is not much change in the shape of the waveform), above all when using linear interpolation to calculate the equation between two points.

Figure 8 shows the difference value of the average recognition rate of the adaptive speech speed adjustment algorithm network on the two databases. Comparing the two libraries, it is found that adaptive speed adjustment algorithm of the network and the difference of the remaining three classifier (that is, the recognition of the increased) in TYUT2.0 database to increase the recognition rate is greater than the EMO-DB database, showing that the proposed adaptive speed adjustment algorithm on TYUT2.0 database

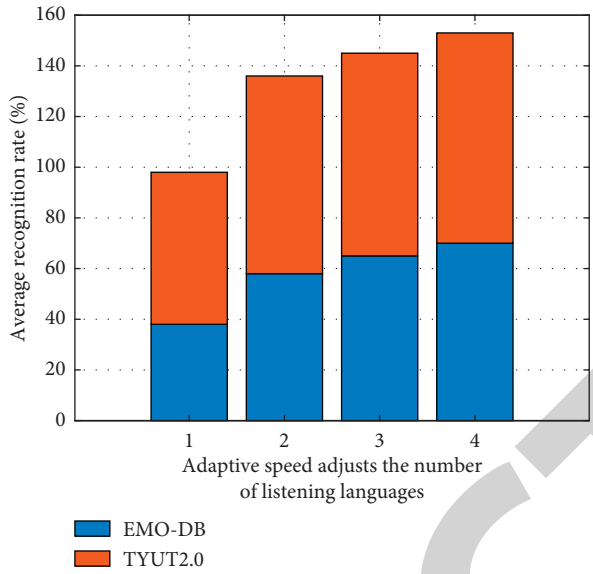

Figure 8: Histogram of average recognition rate of two language databases.

Table 4: Experimental results of the application of the long binary probability to the refined English listening and audio language binary group.

Speed	System	Insert error (%)	Delete the error (%)	Instead of error (%)	Total error rate (%)
Slow	One-variable probability system	5.75	0.16	20.03	25.92
	The refined binary probability system	5.16	0.14	19.45	24.63
Fast	One-variable probability system	0.48	1.43	25.64	27.46
	Refined binary probability system	0.37	1.29	25.42	27.08

than on the EMO-DB database has more advantages. The results show that the adaptive speed adjustment algorithm can better deal with natural speech emotion.

The establishment of the fast algorithm must require the correlation coefficient R BBB 0, but there is a negative correlation coefficient in the statistics, which must have a certain impact on the performance. For this reason, we investigated the negative correlation coefficient (accounting for about 4%) of the binary combination, analysis of its negative reason is that the same English listening audio language in Chinese with different initials when the formation of different syllables, so an English listening audio language binary group in fact corresponding to the binary group of syllables is multiple. In most cases, an English listening audio language binary corresponding to several syllables in the binary group; the former English listening audio language after an English listening audio language influence is roughly the same, so the statistical correlation coefficient is mostly positive and relatively large. But there are also a few exceptions; an English listening audio language duality corresponding to different syllabic duality in the former English listening audio language segment length or the latter English listening audio language to their mean deviation trend is not consistent, resulting in the statistical average; the correlation coefficient is small or negative. In order to overcome the negative correlation coefficient on the performance of the recognition system, the same English listening audio language binary groups corresponding to different syllabic binary groups were separated to find the correlation coefficient, that is, the English listening audio language binary group of correlation coefficient according to different syllables and then refine, so as to avoid the occurrence of negative correlation coefficient.

It can be seen from the experiment that the adaptive algorithm of speech speed is mainly aimed at the situation that the speech speed is far from the average speed, so the experiment is only carried out for the slow speed and the fast speed, and the results are shown in Table 4.

5. Conclusion

In this paper, an adaptive speech speed adjustment algorithm is introduced to detect phonemic attributes. The algorithm takes sentence as unit and uses continuously changing frame length and frame shift interval to normalize the speech speed, so that the adjusted speech speed is consistent with the average speech speed of corpus. In addition, when detecting the speech speed of the test set, an adaptive algorithm based on the length of adjacent speech units is proposed, which achieves a good effect on the experiment of digital string, and the total error rate is relatively reduced at slow speed, and the boundary of phonemes and silent segments are accurately detected; thus, the speech speed is obtained indirectly. After adjusting the speed of speech, the dynamic range of continuous frames in time of phonemic attribute is reduced, and the performance of

HMM based system is improved. The experiment shows that the algorithm can automatically adjust the length of speech unit according to the speed of speech when the corpus deviates from the normal speed, thus reducing the insertion error and deletion error caused by the speed of speech, obtaining more accurate segmentation points, and thus reducing the substitution error, thus improving the performance of the system. What needs to be studied in the future is the adequacy of the adaptive training of speech speed in the continuous speech recognition with large vocabulary.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Consent

Informed consent was obtained from all individual participants included in the study references.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The study was supported by the first-class discipline construction on pedagogy by the Institution of Higher Education of Ningxia (Grant no. NXYLXK2017B11).

References

[1] A. Sayoud, M. Djendi, and A. Guessoum, "A new speech enhancement adaptive algorithm based on fullband-subband MSE switching," *International Journal of Speech Technology*, vol. 22, no. 4, pp. 993–1005, 2019.

[2] H. Phan, L. Hertel, M. Maass, P. Koch, R. Mazur, and A. Mertins, "Improved audio scene classification based on label-tree embeddings and convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1278–1290, 2017.

[3] J. K. Jia, Y. B. Shao, H. Long, and Q. Z. Du, "A natural language sentence analysis algorithm based on word order modifier syntax rules," *Procedia Computer Science*, vol. 166, pp. 496–500, 2020.

[4] J. Li, "Performance evaluation of English part of speech tagging based on multi-feature knowledge algorithm," *Journal of Physics Conference Series*, vol. 1533, pp. 22066–22076, 2020.

[5] F. Chelali and A. Djeradi, "Audiovisual speaker identification based on lip and speech modalities," *The International Arab Journal of Information Technology*, vol. 14, no. 1, pp. 99–110, 2017.

[6] N. Jiang and J. Li, "Adaptive speech enhancement algorithm based on Hilbert-Huang transform," *Ingénierie des Systèmes d'Information*, vol. 24, no. 1, pp. 57–60, 2019.

[7] Z. Song, "English speech recognition based on deep learning with multiple features," *Computing*, vol. 102, no. 99, pp. 1–20, 2020.

[8] M. Kang and J.-M. Kim, "GA-based adaptive window length estimation for highly accurate audio segmentation," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 1, pp. 429–436, 2015.

[9] H. Wang, Y. Gao, and M. Zhou, "Adaptive linear prediction based on compressed sensing algorithm for speech data," *Journal of Information & Computational Science*, vol. 12, no. 11, pp. 245–265, 2015.

[10] C. Zhang, K. Koishida, and J. H. L. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1633–1644, 2018.

[11] Q. Zhang, X. Zhao, and Y. Hu, "A classification retrieval method for encrypted speech based on deep neural network and deep hashing," *IEEE Access*, vol. 8, pp. 202469–202482, 2020.

[12] S. S. Tomic, Z. H. Peric, and J. R. Nikolic, "An algorithm for simple differential speech coding based on backward adaptation technique," *Informatica*, vol. 29, no. 3, pp. 539–553, 2018.

[13] V. Das, A. Kar, and M. Chandra, "Advanced adaptive algorithms for double talk detection in echo cancellers: a technical review," *Advances in Intelligent Systems and Computing*, vol. 328, pp. 297–305, 2015.

[14] F. Huang, T. Lee, W. B. Kleijn, and Y.-Y. Kong, "A method of speech periodicity enhancement using transform-domain signal decomposition," *Speech Communication*, vol. 67, pp. 102–112, 2015.

[15] R. M. Ramli, A. O. A. Noor, and S. Abdul Samad, "Noise cancellation using selectable adaptive algorithm for speech in variable noise environment," *International Journal of Speech Technology*, vol. 20, no. 3, pp. 535–542, 2017.

[16] R. Soleymani, I. W. Selesnick, and D. M. Landsberger, "ALTIS: a new algorithm for adaptive long-term SNR estimation in multi-talker babble," *Computer Speech & Language*, vol. 58, no. 11, pp. 231–246, 2019.

[17] S. Huang and C. Li, "Distributed extreme learning machine for nonlinear learning over network," *Entropy*, vol. 17, no. 2, pp. 818–840, 2015.

[18] G. A. Khan and K. Murthy, "Regularized NLMS adaptive algorithm for noise cancellation in speech signals," *IOSR Journal of Electronics and Communication Engineering*, vol. 11, no. 4, pp. 41–45, 2016.

[19] G. A. Khan and K. Murthy, "Implementation of optimized proportionate adaptive algorithm for acoustic echo cancellation in speech signals," *International Journal of Electronics Engineering Research*, vol. 9, no. 6, pp. 823–830, 2017.

[20] J. Yang, Z. Bian, Y. Zhao, W. Lu, and X. Gao, "Full-reference quality assessment for screen content images based on the concept of global-guidance and local-adjustment," *IEEE Transactions on Broadcasting*, pp. 1–14, 2021.

[21] Y. Li and J. Yang, "Few-shot cotton pest recognition and terminal realization," *Computers and Electronics in Agriculture*, vol. 169, no. 6, Article ID 105240, 2020.

[22] X. Cheng, B. Yang, A. Hedman, T. Olofsson, H. Li, and L. Van Gool, "NIDL: a pilot study of contactless measurement of skin temperature for intelligent building," *Energy and Buildings*, vol. 198, pp. 340–352, 2019.

[23] W. Wang, Z. Gong, J. Ren, F. Xia, Z. Lv, and W. Wei, "Venue topic model—enhanced joint graph modelling for citation recommendation in scholarly big data," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 20, no. 1, pp. 1–15, 2021.

[24] C. Chen, C. Wang, T. Qiu et al., "A robust active safety enhancement strategy with learning mechanism in vehicular networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 12, pp. 5160–5176, 2019.
[25] W. Wei, J. Wu, and C. Zhu, "Special issue on deep learning for natural language processing," *Computing*, vol. 102, no. 3, pp. 601–603, 2020.