Hindawi
Complexity
Volume 2023, Article ID 9892303, 1 page
https://doi.org/10.1155/2023/9892303

Retraction

Retracted: English Phrase Speech Recognition Based on Continuous Speech Recognition Algorithm and Word Tree Constraints

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This article has been retracted by Hindawi following an investigation undertaken by the publisher [1].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] H. Du and H. Duan, "English Phrase Speech Recognition Based on Continuous Speech Recognition Algorithm and Word Tree Constraints," *Complexity*, vol. 2021, Article ID 8482379, 11 pages, 2021.

Hindawi
Complexity
Volume 2021, Article ID 8482379, 11 pages
https://doi.org/10.1155/2021/8482379

Research Article

English Phrase Speech Recognition Based on Continuous Speech Recognition Algorithm and Word Tree Constraints

Haifan Du[1] and Haiwen Duan[2,3]

[1]*Ordos Vocational College, Ordos 017000, China*
[2]*China University of Mining and Technology, Xuzhou 221116, China*
[3]*Inner Mongolia Yitai Group Co., Ltd., Ordos 017000, China*

Correspondence should be addressed to Haiwen Duan; ds20070137p32mb@cumt.edu.cn

Received 9 April 2021; Revised 29 April 2021; Accepted 14 May 2021; Published 25 May 2021

Academic Editor: Zhihan Lv

This paper combines domestic and international research results to analyze and study the difference between the attribute features of English phrase speech and noise to enhance the short-time energy, which is used to improve the threshold judgment sensitivity; noise addition to the discrepancy data set is used to enhance the recognition robustness. The backpropagation algorithm is improved to constrain the range of weight variation, avoid oscillation phenomenon, and shorten the training time. In the real English phrase sound recognition system, there are problems such as massive training data and low training efficiency caused by the super large-scale model parameters of the convolutional neural network. To address these problems, the NWBP algorithm is based on the oscillation phenomenon that tends to occur when searching for the minimum error value in the late training period of the network parameters, using the K-MEANS algorithm to obtain the seed nodes that approach the minimal error value, and using the boundary value rule to reduce the range of weight change to reduce the oscillation phenomenon so that the network error converges as soon as possible and improve the training efficiency. Through simulation experiments, the NWBP algorithm improves the degree of fitting and convergence speed in the training of complex convolutional neural networks compared with other algorithms, reduces the redundant computation, and shortens the training time to a certain extent, and the algorithm has the advantage of accelerating the convergence of the network compared with simple networks. The word tree constraint and its efficient storage structure are introduced, which improves the storage efficiency of the word tree constraint and the retrieval efficiency in the English phrase recognition search.

1. Introduction

Voice, the sound of language, is an important vehicle for human communication. The traditional keyboard input can hardly meet the requirements of today's consumers for the portability and efficiency of electronic products. To make machines understand human language and realize efficient and convenient communication between humans and machines has become a hot topic of research today. The research of speech recognition needs to involve many professional disciplines, including acoustics, digital signal processing, pattern recognition, probability theory and information theory, vocal mechanism and auditory mechanism, artificial intelligence, and many other disciplines [1]. Speech recognition is likened to the "auditory system of a machine," and the ideal result of recognition is that the machine understands what people say. In a variety of complex environments, the correct acquisition of speech information content by various speech processing algorithms and the execution of various intentions of the speaker based on semantic information are important means to achieve natural interaction between humans and machines [2]. Speech recognition is gradually integrated into people's lives with its broad social benefits and application prospects, such as smartphones, remote control of home appliances, intelligent robots, voice navigation, and other products [3].

There are still many difficulties in speech recognition that need to be solved, such as the poor robustness of feature parameters and the lack of accuracy of base tone cycle extraction.

Although speech recognition technology has made significant progress, the performance of speech recognition technology in a nonideal environment needs to be further improved. The difficulty of speech recognition in a nonideal environment lies in the fact that the speech signal collected in this situation is too much affected by perturbation factors, among which the perturbation factors are concentrated in the following aspects: (1) individual differences in the speech signal, including individual characteristics such as speaker's accent, emotion, speech speed, coarticulation, and pause [4]. (2) Uncertainty brought by speech signal collection equipment, due to the performance difference of collection equipment, even the speech acquisition signal of the same individual in the same environment can have great differences [5]. (3) The background noise in the nonideal environment is uncontrollable, which may be noisy hawking, whistling gale, or soothing music. The distribution of the noise is random, and it is difficult to strip the noise from the speech signal. Phrase recognition research plays a pivotal and supportive role in many fields, for example, semantic analysis, semantic disambiguation, automatic digest, information retrieval, and information extraction; and it occupies an indispensable position in multilingual information retrieval systems, human-computer dialogue systems, word disambiguation, lexicon compilation and updating, automatic text classification, and search engines [6].

The purpose of this paper is to study the aspects of improving endpoint detection accuracy, reducing data set variability, improving acoustic model training algorithms, and designing and implementing system prototypes in the process of speech recognition, to improve the efficiency and accuracy of speech recognition. Section 1 introduces the research value of speech recognition, clarifies the necessity of improving the accuracy and efficiency of speech recognition, focuses on the current status of domestic and international research on improving the accuracy of speech recognition, and finally makes a description of the main research content of this paper. Section 2 introduces the general acoustic model and the problems encountered in the process of parameter training, then describes the relevant technologies needed in the preprocessing stage of the speech signal, and conducts an in-depth study on the working principle of convolutional neural network and parameter training algorithm, to prepare the theoretical foundation for the subsequent research work. Section 3 proposes a reduced weight range backpropagation (NWBP) algorithm, and the training process also uses the error backpropagation algorithm, while the search method based on word tree constraint for English phrase speech recognition is studied to improve the search efficiency of the speech recognition algorithm. Finally, the prototype system of speech recognition is realized through system structure construction, main module design, and module integration. Section 4 verifies the applicability of the system to specific environments and the effectiveness of the algorithm by testing the accuracy of speech data in different environments. Section 5 summarizes the research contents and results of this paper and proposes the next work to be completed and new research directions because of the difficulties and shortcomings encountered in the research process. (1) The performance of our recognition algorithm is very stable, (2) the accuracy of our results is improved by about 10.5% compared to other studies, and (3) our research can be applied to real-life applications.

2. Related Works

The proposal of the multilayer perceptron model marks the beginning of the machine learning era. Speech recognition technology has been able to be studied and developed more deeply, and the introduction of artificial neural networks and the combination of models have brought speech recognition technology to a new stage. Zhou et al. took English verb phrases as the object and realized verb phrase recognition by lexical annotation, named entity recognition, and rule constraints. Its process is relatively tedious [7]. Zerari et al. identified labelled basic noun phrases by considering the internal structural features of the phrases, using the corresponding lexical sequences as rules and then performing clipping of the lexical sequences to obtain rule sets. However, the accuracy of recognition needs to be improved [8]. Cui et al. selected relevant samples from many stores and used some conventional methods for feature extraction of spectral and rhyme types for a few selected types [9]. Applying them to Recurrent Neural Network (RNN) and Fully Focused Time Delay Neural Network (FFTDNN) structures to evaluate their performance in recognizing mood, dialect, speaker, and gender differences in dialectal Assamese languages, it was found that machine learning-based (machine learning, ML) sentence extraction technique along with RNN using a composite feature model as a classifier outperformed other methods in terms of recognition rate and computational efficiency under several background noise conditions [10]. Shen et al. proposed an efficient speech data selection technique to improve recognition data speed. This technique selects speech data that is acceptable for speech recognition applications and reduces the time required to compute confidence scores by traditional confidence measurement techniques [11]. The technique is based on acoustic likelihood values and can quickly select speech data with high a priori confidence. Experiments show that the proposed confidence estimation technique is more than 50 times faster than traditional postconfidence measures while providing equivalent data selection performance for speech recognition and verbal document retrieval [12].

The phrase objects studied in this paper are fixed phrases or quasi-fixed phrases, which are generally characterized by stable structure and high cooccurrence frequency, and are recognized and semantically analyzed in syntactic analysis. The algorithm is applied to the speech system of the continuous speech recognition model, and the superiority of the algorithm is compared and analyzed through the training and recognition of the feature parameters. Finally, the software platform of MATLAB is used to design a set of the user interface (GUI) of the speech recognition system based

on a deep learning model to show the environment of speech recognition preprocessing, improved feature extraction parameters, speech sample training, and speech recognition in the form of operation interface, which is convenient for subsequent data analysis and processing.

3. English Phrase Speech Recognition Study

3.1. Research on Continuous Speech Recognition Algorithm Based on Deep Learning. Most speech recognition systems are built based on the Hidden Markov Model, which is a pattern recognition method that can simulate the temporality of speech signals. Usually, the system performs feature extraction on the original speech signal and converts it into a feature sequence [13]. The purpose of speech recognition is to combine the acoustic model and the language model to finally recognize the corresponding word sequence H according to the maximum a posteriori probability criterion, whose mathematical expression is shown in the following equation:

$$H = \left(\sum_{i=1}^{m} \arg(h_i)\right) \cdot \max\left(P\left(\frac{M^t}{H}\right) \cdot P(H)\right). \quad (1)$$

In practice, in the flat region of the error surface, the error gradient information is extremely small, making the network converge slowly, or even the network cannot converge, which in this region requires a larger learning rate; otherwise, it falls into the local minimum. In the region where the error surface changes a lot, the amount of change in the weights is relatively large, using the original learning rate will cause oscillations due to repeated changes in the weights around the very small values, and the algorithm is not easy to converge. In this case, the learning rate should be reduced [14]. The trend of the error surface change is judged by the mean square error change w of two consecutive iterations, and the threshold parameter x is set to limit the curvature change, which is calculated as shown infd2

$$\begin{cases} \beta(x) \cdot m, & m \in [0,1], \quad |w_x - w_{x-1}| \geq x, \\ \beta(x) \cdot n, & n > 1, \quad |w_{x+1} - w_x| < x. \end{cases} \quad (2)$$

In the previous equation, p and q are set values to constrain the variation of the learning rate β and thus adapt to the error of different trends, improving efficiency while reducing oscillations and local minima.

Based on the variable step-size method to further improve the learning rate variable so that it can change its size adaptively with the change of training progress, from the perspective of how fast or slow the model learns, an improved variable learning rate algorithm was proposed [15]. For the BP algorithm, the learning rate cannot adapt to the sudden change of the error surface and cannot achieve a good smooth transition, the IVLBP algorithm can make the mean square error converge smoothly along the gradient direction when the curvature changes abruptly by setting two threshold parameters to measure the increase and decrease of the mean square error, and it is easier to reach the global minimum point; the algorithm is shown in

$$\begin{Bmatrix} H^m(x+1) = H^{m+1}(x) \cdot H^{m-1}(x-1) & H^m(x+1) = \beta H^{m+1}(x) + \alpha H^{m-1}(x) \\ \forall E > \zeta^2 & \zeta^2 \geq \forall E \\ x = \alpha x, \quad \alpha \in [0,1] & x = \sigma x, \quad x > 1 \end{Bmatrix}. \quad (3)$$

The K-means algorithm is used to find the range where the error minima may occur. In the middle and later stages of training, the network is learned using the IVLBP algorithm based on the results of the weight changes calculated in the most recent iterations to find the initial seed node that approximates the error minima, while the results of the weights calculated in the iterations are recorded. In the convolutional neural network, the backpropagation process of parameter training is divided into convolutional layer propagation and sampling layer propagation, if the next layer of the convolutional layer is the sampling layer, then the next layer of the sampling layer is the convolutional layer, so the residual calculation of the convolutional layer is one-to-one nonoverlapping sampling, and the residual calculation of the feature map of the l^{st} layer is shown in

$$\begin{aligned} \eta_i^j &= \frac{\sum_k^j \forall k}{\sum_k^j \forall k} \\ &= \omega_i^j\left(h\left(\beta_i^j\right)\right) \otimes \delta up\left(\delta_i^j\right). \end{aligned} \quad (4)$$

The derivative of the bias value u in the convolutional layer and the derivative of the bias parameter v in the convolutional kernel are calculated below:

$$\frac{\forall H}{\forall k_{ij}^l} = \sum_u^v \sum_i^j \delta_i^j h(uv) H_{ij}^l(\delta)_{uv}. \quad (5)$$

The $p+2$ calculation is performed by the IVLBP algorithm, and during the network learning process, the weighting coefficients are increased or decreased by adaptively changing the learning rate according to the curvature

of error change, so that the error function M decreases in the direction of the fastest decrease, and the coefficient update is calculated as shown in

$$\eta_p^q = M_p^q \cdot \delta_i^j \cdot \frac{1}{M_i^p} \cdot \left(q_i^p - p_j^p\right). \tag{6}$$

A pure speech waveform is used for the experiment, such as the speech content of "1, 2, 3," and the noise is added later, the pure speech and the noisy speech are shown in Figures 1(a) and 1(b), and the endpoint detection of the noisy speech is performed by using the cosine angle value of the autocorrelation function; the time and the cosine angle value of the autocorrelation function of the speech segment are obtained as shown in Figure 1(c). The solid line indicates the starting point of the speech, the dashed line indicates the endpoint of the speech, and the effective speech segment is indicated between the starting point and the endpoint.

A higher threshold T2 is selected for coarse judgment according to the variation of short-time energy of speech, and the speech segment is above the threshold T2, while the speech starts and endpoints are located outside the time point corresponding to the intersection of this threshold and the short-time energy envelope point [16]. The lower threshold T1 is determined on the average energy, and the point that intersects with the horizontal line T1 is found from the outer circle of the intersection point, respectively, which is the start or endpoint of the speech, and the speech is processed according to the short-time energy obtained above combined with the endpoint detection method described above:

$$F(x) = \beta \cdot \cos(x) \cdot \frac{\sum_i^m M_H(m) M_i(m)}{\min\left(\left\| M_H(m) M_i(m) \right\|\right)}. \tag{7}$$

The enhanced speech short-time energy can better highlight the signal characteristics of the speech segment without increasing the speech energy of the noisy segment when it passes through the double threshold detection. The processed noisy speech can be better distinguished from the speech segment by the threshold setting to achieve more accurate speech endpoint detection.

If the mean squared error (over the entire training set) weights increase after the update and exceed a certain set percentage (typical values are 1% to 5%), the weight update is cancelled, the learning speed is multiplied by a factor p ($0 < p < 1$), and the momentum coefficient y (if any) is set to 0. If the squared error decreases after the weight update, the weight update is accepted, and the learning speed is multiplied by a factor $n > 1$. If y is set to 0, it reverts to its previous value. If the squared error increases by less than, then the weight update is accepted, but the learning rate remains unchanged. If y has been set to 0 in the past, it reverts to its previous value.

3.2. Research on English Phrase Speech Recognition Based on Word Tree Constraints. Phrase (phrase) recognition under rule constraints is mainly summarized manually or semi-artificially to find out what are forms of the close union between words and phrases, which are generally expressed formally using class connections (Table 1). After the class connection is established, the rule constraint is carried out based on the generative rules, combined with the contextual information, and in the implementation, attention is paid to the order and optimization issues between rules and rules, which cannot contradict or conflict with each other. The rule constraint is identified and extracted by simulating the matching of the dictionary, and if the matching is successful, the phrase is output as true; if the matching fails, the phrase is output as false. Here, the simulated dictionary can be established by manual collation, or by some algorithm implementation for word classification extraction and collation, and then manually screened again in the way. Batch is 10 means you calculate 10 data samples in one batch; each data sample should be in the form of $m * n$ matrix, so your actual input is $10 * m * n$ 3D matrix, where m is time_step and n is feature_num; after convolution and full connection, and your output is $10 * p * q$, where q means m after convolution operation. q is num_classes, the last fully-connected layer is the output layer, used for classification, and num_classes is the number of categories corresponding to the data set.

The addition of the lexicon is an auxiliary method to the rule constraint. The determination of the lexicon weights is feasible either by testing the statistical selection of the optimal value or by a related algorithm, such as the EM algorithm. Naturally, the larger the lexicon, the better, but we are also aware that language changes over time and that lexicons are a difficult task to organize. Besides, the data contained in lexicons is generally very stable and tightly bound, and the lexicons are disadvantaged in extracting and identifying quasi-phrase data or free phrases. Matching recognition by dictionaries is only possible for small-scale and stable linguistic phenomena.

We all know that a power signal is defined as a signal with finite average power and infinite energy, while an energy signal is defined as a signal with average power equal to zero and finite energy. Such a classification is useful for our discussion of comparing analog-digital signals. Because the analog signal waveform lasts infinitely long, without splitting or adding a time window, and its energy is infinite, it cannot be described by energy, which means that power is a better parameter. In contrast, waveforms of time duration Ts between code elements are used in digital systems to send and receive code elements. The power is naturally zero in the whole-time axis, so it cannot be described by it, and the signal is generally measured in the time window. Therefore, energy is a better descriptive parameter.

The deep learning analysis comes first, and the rule constraints are used as the postprocessing part to assist in recognition, in which the lexical factor plays an appropriate supporting role as an independent variable factor in the binding degree calculation [17]. First, input the text to be tested and enter the preprocessing part, which mainly includes language judgment, word division, lexical annotation, delimitation, and other branches; then, traverse each utterance in the text for phrase candidate string extraction, and output the utterance without phrase information if there is no phrase

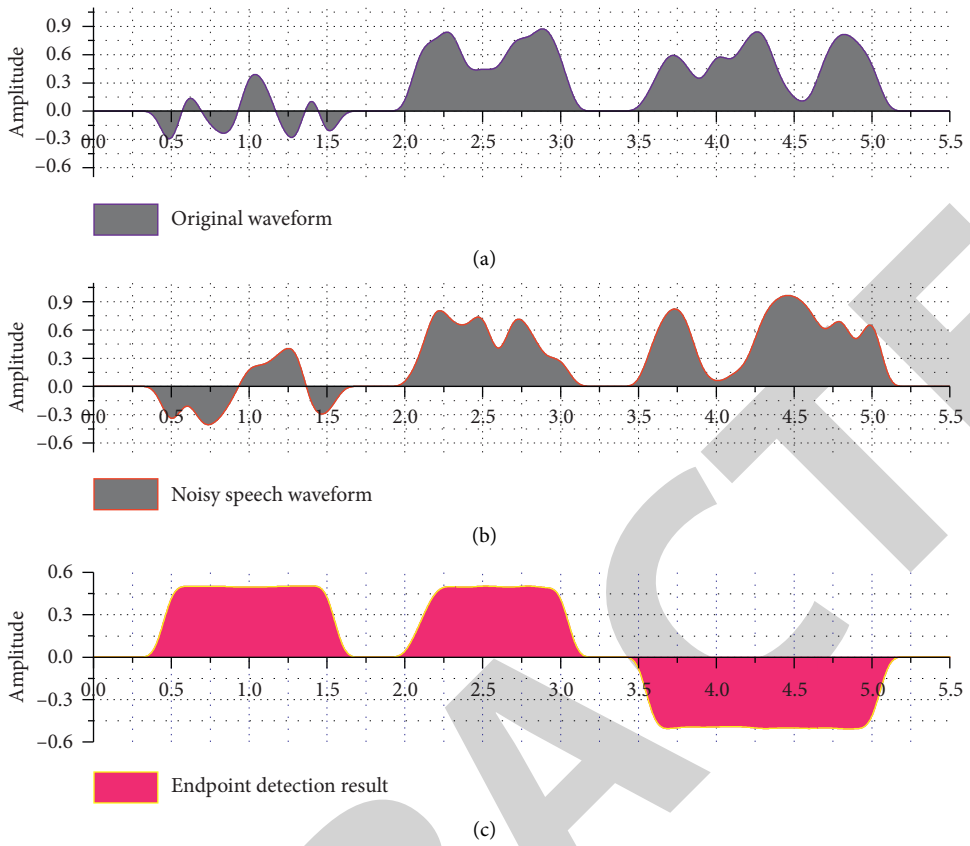

Figure 1: Continuous speech endpoint detection.

Table 1: Class connection methods.

Serial number	Class connection form	Detailed explanation
1	DET + N + N	Qualifier + noun + noun
2	V + N	Verb + noun
3	N + PART	Noun + particle
4	V + PART + PREP	Verb + particle + preposition
5	PREP + N	Preposition + noun
6	N + V	Noun + verb

candidate string; if there is a phrase candidate string, the binding degree is calculated and analyzed, and naturally, the binding degree can be added to the simulated phrase lexicon as needed. Naturally, the combination degree can be added to the simulated phrase lexicon as needed to assist the judgment. Next, we enter the postprocessing part, mainly the addition of rule constraints, class conjunction, and contextual grammar constraints, to further compensate for some shortcomings of the probabilistic analysis method [18]. After the sentence is judged, the text statement is analyzed to see if the traversal is completed, and if it is not, the statement traversal part is entered; if the traversal is completed, it is finished. Through the fusion of the above methods, the system finally achieves high accuracy and robustness in recognizing phrases.

3.3. English Phrase Speech Recognition System Implementation Research. The overall framework of the speech recognition system is shown in Figure 2, which mainly includes four modules, namely, feature extraction, acoustic model, language model, and decoder. The training process is to take the original speech signal as input for feature extraction and transform the speech signal from the time domain to the frequency domain; then, the extracted acoustic features are trained statistically to get the acoustic model, and the text is trained to get the language model. In the recognition process, the features of the speech signal are also extracted, and after the representation of the trained acoustic model, the recognition result is obtained by combining the language model and the dictionary through the decoder.

In this paper, we choose Tensor flow to build an English speech recognition system. Tensor flow integrates a variety of function packages required in the field of deep learning and encapsulates some complex functions and functions to simplify the details of model building. In Tensor flow, nodes are used to represent different function operations, and each edge in the structure represents the data interaction between the operation nodes. Tensor represents the data transferred

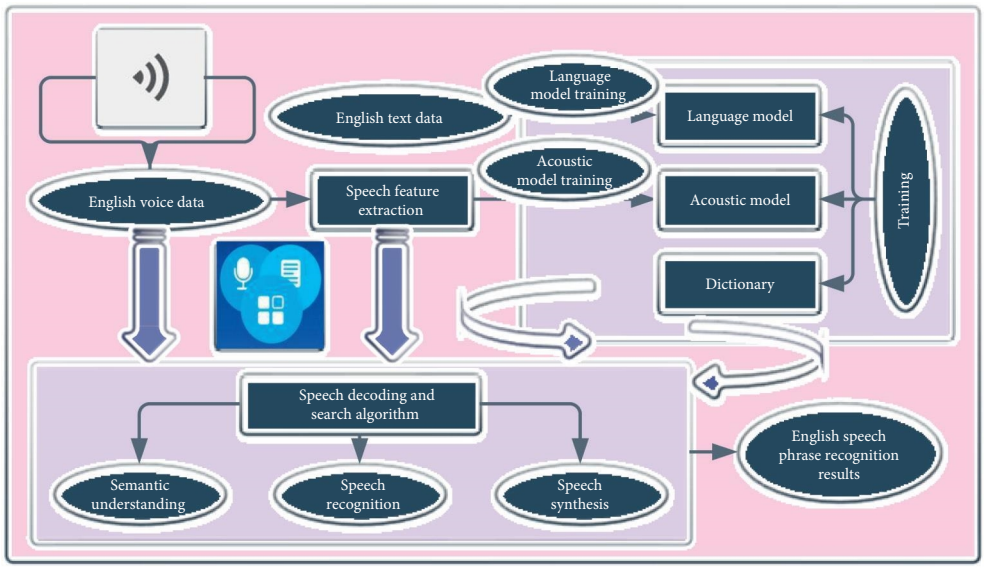

Figure 2: The framework of the speech recognition system.

between different nodes, usually a multidimensional matrix or vector; Flow is the information flow, which can be understood as data information in a form of flow into each node through the whole operation graph [19].

For the English phrase database, which is divided into the training set and test set, the total duration of the speech library is over 50 hours. The training set is labelled as an ABC group with 250 sentences each, and the test set is group D. The sampling frequency is 16 kHz and the sample size is 16 bits. The speech recognition model is trained and learned by the ABC group, and the recognition effect is finally characterized by the D group. The speech composition of English phrases is shown in Table 2.

Table 2: Phonetic composition of the data set.

Data content	Data unit	Training set	Test set
Speaker	Person	321	13
Female	Person	36	2
Male	Person	24	9
Age range	Year	19–61	18–53
Number of voices	Article	11457	2643
Time-consuming	Hour	28.14	7.12

4. Results and Analysis

4.1. Analysis of Continuous Speech Algorithm. In CNN, the stochastic gradient descent algorithm (SGD) is usually used to train the model in a supervised manner. The training process is divided into two stages: forward propagation and backward propagation. In the forward propagation process, the speech features pass through each layer of the CNN and obtain an output value to calculate the error for back-propagation. In the backpropagation stage, the weights and biases of each layer need to be calculated and adjusted by the errors. In this paper, we set the number of iterations to 16, the initial learning rate to 0.01, and the minibatch size to 256 and validate the CNN model and CTC-CNN model for the English speech recognition system. Compared with the CNN acoustic model, the CTC-CNN acoustic model can get a lower word error rate and better recognition effect in English speech recognition. The acoustic model word error rate is shown in Figure 3.

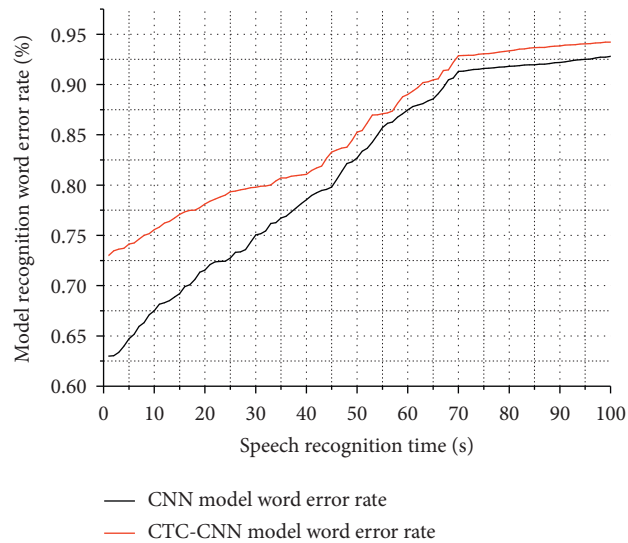

Figure 3: The acoustic model word error rate.

To test the robustness and practicality of the algorithm, a total of 200 English speech phrases were recorded in a quiet environment by 100 people of each gender, all of whom spoke Mandarin daily and were of different ages. The phonetic, rhyme, and tone distributions of the recorded syllables were generally consistent with the acoustic statistics of Mandarin. Gaussian white noise with signal-to-noise ratios of −10 dB, 0 dB, and 10 dB was added to the 200 English speech phrases, and the bass tones were detected by each of the three methods mentioned above. The average

amplitude difference function of the pure speech signal combined with manual correction is used as the reference standard, and the detection result of more than 10 artifacts is considered as one false detection. From the results of Figure 4, we can see that as the signal-to-noise ratio decreases and the correct rate of this algorithm has been maintained above 90.05%, which is much higher than the other two algorithms and has the best robustness.

4.2. Analysis of English Phrase Speech Recognition System. Figure 5 shows a comparison of the effect of different MFCC parameter extraction algorithms for a frame of the turbid speech signal in the English speech segment. Figure 5(a) shows the 12-dimensional static MFCC feature parameters extracted using the traditional feature extraction algorithm. Although the computational complexity is reduced due to the numerical limitations of the feature parameters, the dynamic characteristics between adjacent frames of the speech signal are not well-reflected. Figure 5(b) shows the 24-dimensional dynamic MFCC feature parameters composed of traditional MFCC parameters combined with MFCC first-order difference. Compared with the previous traditional algorithm, the feature values contain part of the dynamic information of speech, and the feature value information coverage is more comprehensive, but the 12-dimensional static MFCC parameters are not smooth enough, and the value fluctuation of the latter 12-dimensional dynamic parameters is small. The discrimination effect is not good.

The results of the feature extraction algorithm in this paper are shown in Figure 5(c) and the multilayer IMF decomposition of the speech signal using EMD decomposition; it can be seen that the MFCC parameters obtained after empirical mode decomposition differ from those in Figure 5(a) and Figure 5(b) in the first 24 dimensions; the first 12-dimensional static MFCC feature parameters of the improved algorithm are more smooth compared to Figure 5(a) and Figure 5(b), while the first-order difference of the dynamic MFCC in the last 12 dimensions has greater variance fluctuations compared with the data in Figure 5(b); and the dynamic characteristics of the data are more obvious. Moreover, although the use of the fundamental period-related parameters to expand the dimensionality of the feature parameters will increase the computational effort, it can incorporate the intonation information of speech and greatly improve the recognition accuracy, and the advantages outweigh the disadvantages.

Figure 6 shows that the average recognition rate of the extraction algorithm in this paper is 4–8 percentage points higher than that of the MFCC feature parameter extraction algorithm, which is widely used at present, under various signal-to-noise conditions. The recognition rate of three or four-word names is higher than that of two-word urban names because the endpoint-detected feature parameter models are easier to distinguish between speech segments. However, the reduction of the final recognition rate of the improved MFCC combined feature parameter extraction algorithm in this paper is smaller than that of the former,

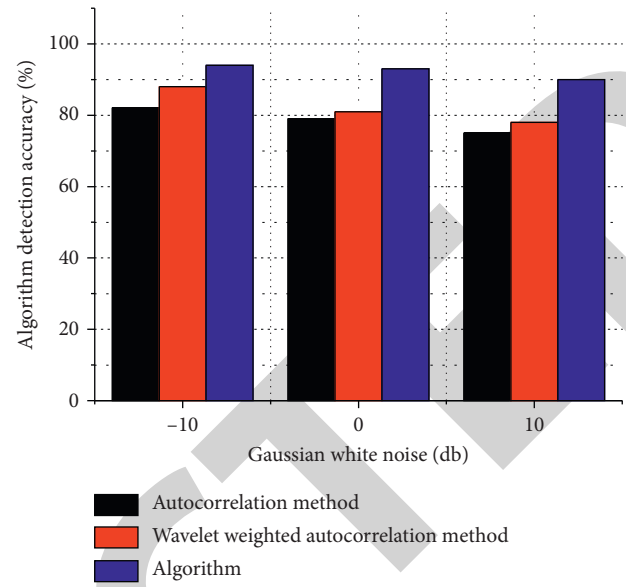

Figure 4: Comparison of the correct rate of base-tone cycle detection of the three methods.

and the robustness is the strongest. In summary, this improved MFCC feature extraction algorithm can be applied to the non-person-specific small vocabulary isolated word recognition system of HMM instead of the traditional MFCC feature parameter extraction algorithm to obtain a higher recognition rate and stronger robustness.

4.3. Speech Recognition Practice Analysis. To investigate the effect of downsampling on the recognition performance aspect, the network structure of the CNN in the acoustic model needs to be fixed, and the results generated by downsampling at different scales are compared in terms of the number of parameters of the model for extracting BN features and the recognition correctness on the final test set, given that the CNN structure is identical. The performance of different downsampling widths s is compared for the expanded stitching part of the frames in the tandem DNN model. The downsampling widths s are chosen as 1, 2, and 3, where $s=1$ means that the expanded stitching part of the frame is not downsampled, and the experimental results are shown in Figure 7.

From Figure 7, (1) the model corresponding to a downsampling width of 3 has the best recognition accuracy when the number of layers of the tandem DNN model used to extract BN features is 4; (2) the model corresponding to a downsampling width of 2 has the highest recognition accuracy when the number of layers of the tandem DNN model used to extract BN features is 5; (3) the number of parameters of the model decreases as the downsampling size increases of the model. The above shows that the same downsampling width has different effects on the model when the number of layers of the tandem DNN model used to extract BN features varies: the recognition accuracy may be higher or lower compared to the model with no downsampling. Although the effect of downsampling is not always

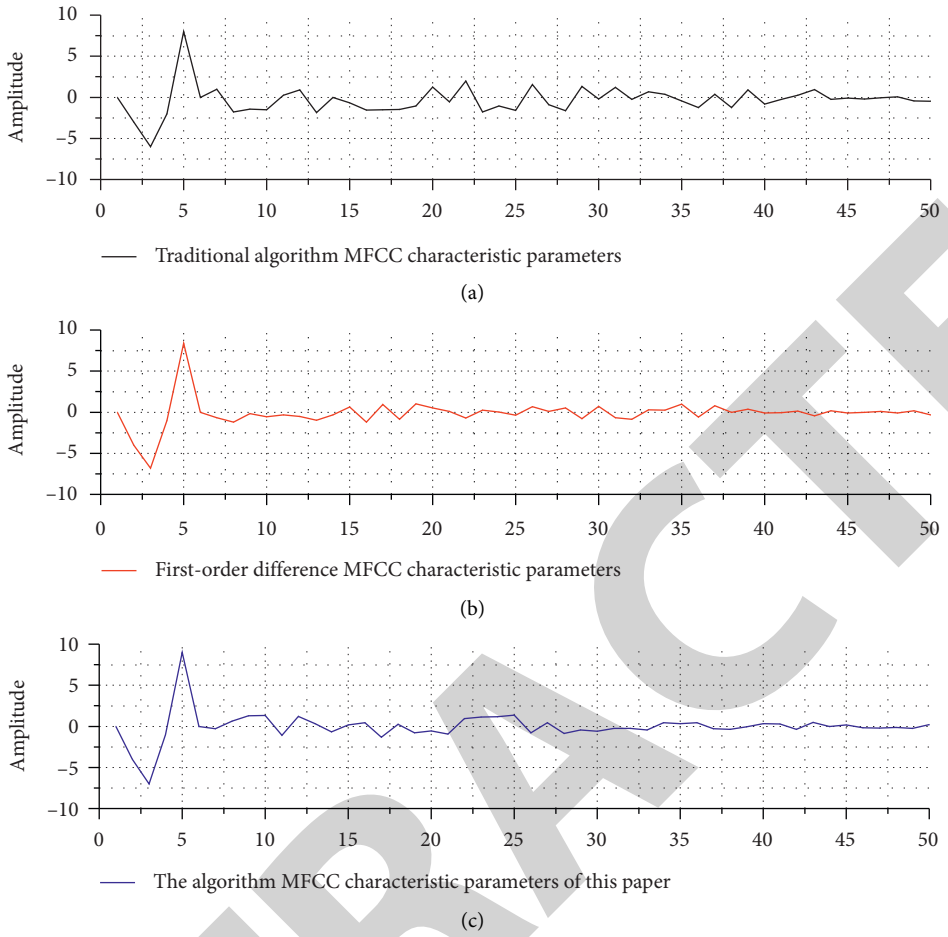

Figure 5: Comparison of three MFCC parameter data in one frame.

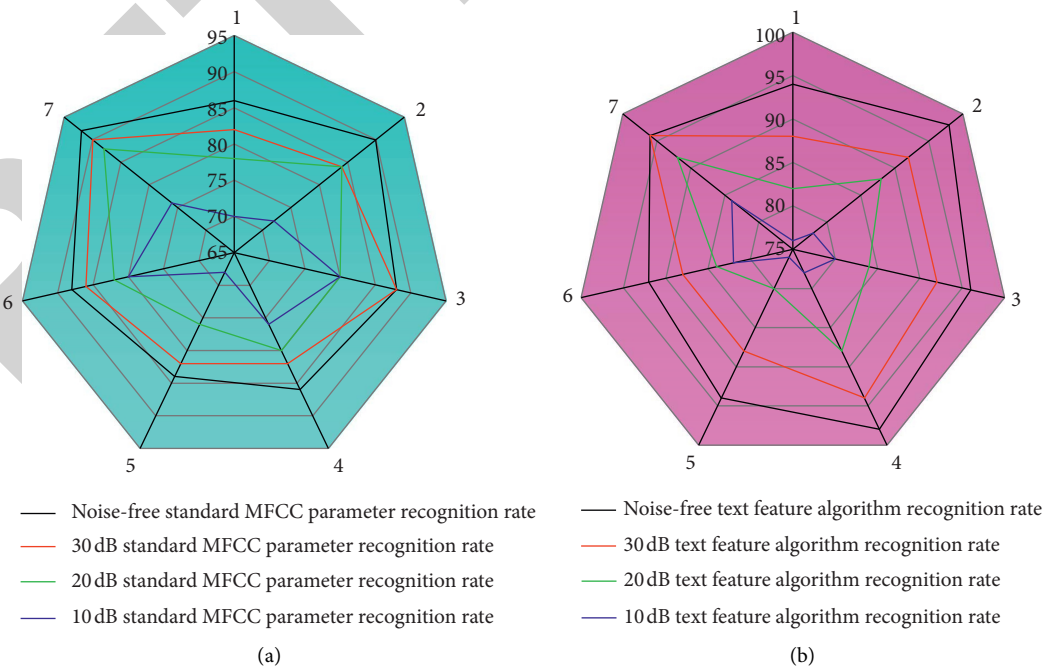

Figure 6: System identification results: (a) the first type and (b) the second type.

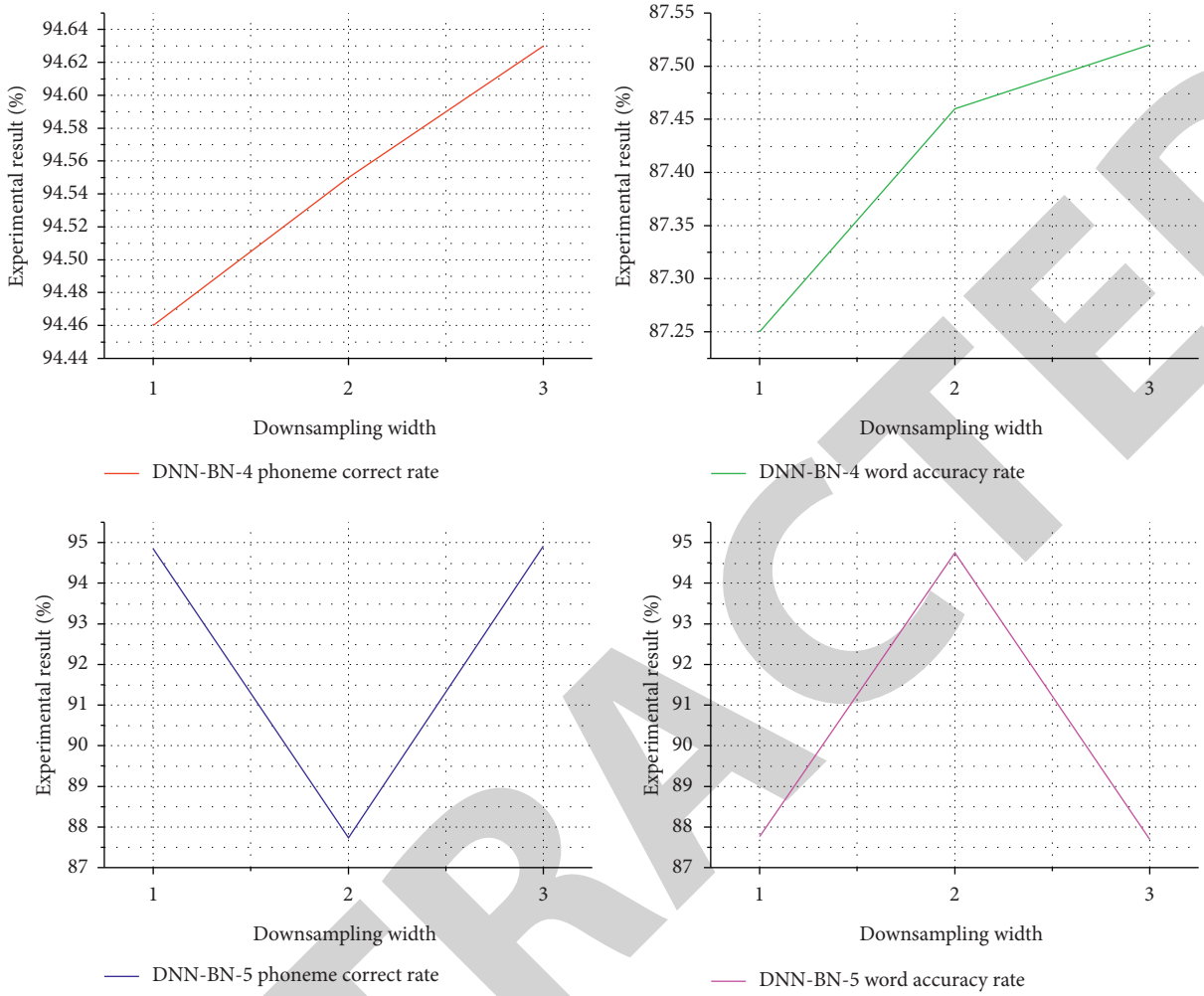

Figure 7: Performance comparison of downsampling.

to improve the recognition performance, there is always a suitable downsampling width to improve the recognition performance of the model when the number of layers of the tandem DNN model used to extract BN features is assumed to be fixed. At the same time, downsampling reduces the number of parameters of the model, which makes the model complexity lower, and therefore, according to Occam's razor principle, the model after downsampling is better than the model without downsampling in model selection. Multicarrier is the superposition of time-domain modulation of multiple modulated symbols, then, generally at the transmitter side, it is necessary to do the normalization of the transmit signal; this normalization is the normalization of the energy of the symbols, because, in the subcarrier, the guide frequency is the energy component, the magnitude of the power of the guide frequency, which directly affects the performance of the receiver.

To investigate the effect of the combination of pretraining and downsampling on the recognition correct rate, the network model for this paper is optimized using pretraining and downsampling, and the recognition correct rate on the final test set is used as an indicator to compare the results generated by the models optimized by different methods. Among them, the optimal width of downsampling is 2. The experimental results are shown in Figure 8. For a deterministic model, both pretraining and downsampling optimization can improve the recognition performance of the network. Pretraining boosts the effect stronger than downsampling. Optimization of the model with both methods will yield the optimal recognition correctness; optimization of the BN feature extraction network with a combination of both will substantially improve the model performance.

In this paper, we analyze the effects of different acoustic models on speech recognition accuracy and further verify the advantages of deep convolutional neural networks with better performance in the English speech recognition process. The experimental results also then show that the end-to-end structure and the application of the mahout activation function improve the performance of the convolutional neural network compared with the traditional convolutional neural network model [20–25]. To address the problem of data sparsity, this paper improves the data sparsity due to the corpus size problem by introducing word similarity calculation in the overall scope. And by data smoothing technique, in the local scope, the excessive bias of statistics calculation

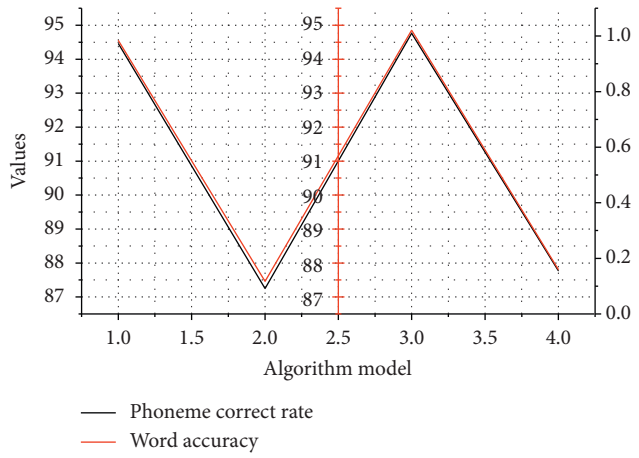

Figure 8: Performance comparison between pretraining and downsampling.

due to local data sparsity is solved because when the frequency of statistics is less than a certain threshold, a large error will occur when the general statistics are used to determine the recognition. Then, by combining the two for use, various problems caused by the natural defects of the corpus itself can be effectively mitigated.

5. Conclusion

The purpose of this paper is to study the aspects of the speech recognition process such as improving the endpoint detection accuracy, reducing the variability of data sets, improving the acoustic model training algorithm, and designing and implementing the prototype system, to improve the efficiency and accuracy of speech recognition. Since speech recognition is mainly applied in the field of maternity and health care, which makes the speech with scene specificity, there will be noise differences between the training set and the speech in the actual application environment. To reduce the difference between the two in terms of background noise, the background noise with environment specificity is added to the training set according to the inverse angle of spectral subtraction, and the noise-containing and noise-free speech are mixed after training processing, which doubles the corpus while enhancing the model robustness of speech recognition in specific environments is improved, and the robustness of the model for noisy speech recognition is improved. Through simulation experiments, the convergence of the NWBP algorithm in the training process of complex convolutional neural network weights is improved compared with the improved variable learning rate backpropagation algorithm, which reduces the redundant computation and shortens the training time to a certain extent, and the algorithm has the advantage of accelerating the convergence of the network compared with the simple network. In this paper, we propose a backpropagation method that incorporates variable learning rate and reduces the range of the distribution of the minima, allowing complex networks to better approximate the error minima, but we need to study how to optimize the algorithm to improve the utilization of storage space to accommodate a large amount of intermediate data during the experiment and how to obtain the optimal seed points in the process of obtaining the optimal seed points by increasing the number of nodes, reasonably adjusting the error distance and the weight of the impact of nodes with alternating errors on the seed points.

Data Availability

The 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.

References

[1] L. Dong, Q. Guo, and W. Wu, "Speech corpora subset selection based on time-continuous utterances features," *Journal of Combinatorial Optimization*, vol. 37, no. 4, pp. 1237–1248, 2019.

[2] S. Bhatt, A. Jain, and A. Dev, "Syllable based Hindi speech recognition," *Journal of Information and Optimization Sciences*, vol. 41, no. 6, pp. 1333–1351, 2020.

[3] Y.-H. Tu, J. Du, and C.-H. Lee, "Speech enhancement based on teacher-student deep learning using improved speech presence probability for noise-robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2080–2091, 2019.

[4] R. Masumura, T. Asami, T. Oba, S. Sakauchi, and A. Ito, "Latent words recurrent neural network language models for automatic speech recognition," *IEICE Transactions on Information and Systems*, vol. E102.D, no. 12, pp. 2557–2567, 2019.

[5] M. Walid, B. Souha, and C. Adnen, "Speech recognition system based on discrete wave atoms transform partial noisy environment," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 466–472, 2019.

[6] S. Singhal, V. Passricha, P. Sharma, and R. K. Aggarwal, "Multi-level region-of-interest CNNs for end to end speech recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 11, pp. 4615–4624, 2019.

[7] L. Zhou, S. Lu, Q. Zhong, Y. Chen, Y. Tang, and Y. Zhou, "Binaural speech separation algorithm based on long and short time memory networks," *Computers, Materials & Continua*, vol. 63, no. 3, pp. 1373–1386, 2020.

[8] N. Zerari, S. Abdelhamid, H. Bouzgou, and C. Raymond, "Bidirectional deep architecture for Arabic speech recognition," *Open Computer Science*, vol. 9, no. 1, pp. 92–102, 2019.

[9] X. Cui, W. Zhang, U. Finkler, G. Saon, M. Picheny, and D. Kung, "Distributed training of deep neural network acoustic models for automatic speech recognition: a comparison of current training strategies," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 39–49, 2020.

[10] N. A. Abu Bakar, R. Ahmad, and A. Sarlan, "Reading fluency evaluation for malaysian primary school children using feature extraction techniques in speech recognition," *Psychology and Education Journal*, vol. 57, no. 9, pp. 478–491, 2020.

[11] Y. Shen, Y. Mai, X. Shen, W. Ding, and M. Guo, "Jointly part-of-speech tagging and semantic role labeling using auxiliary

deep neural network model," *Computers, Materials & Continua*, vol. 65, no. 1, pp. 529–541, 2020.

[12] P. Mittal and N. Singh, "Development and analysis of Punjabi ASR system for mobile phones under different acoustic models," *International Journal of Speech Technology*, vol. 22, no. 1, pp. 219–230, 2019.

[13] B. A. Al-Qatab, M. B. Mustafa, S. S. Salim, and A. A. Sani, "Determining the adaptation data saturation of ASR systems for dysarthric speakers," *International Journal of Speech Technology*, vol. 24, no. 1, pp. 183–192, 2021.

[14] S. Mishra and M. Chandra, "Wavelet-based power normalized spectrum for Hindi phoneme classification," *Circuits, Systems, and Signal Processing*, vol. 38, no. 11, pp. 5149–5168, 2019.

[15] K. Mazur and I. Tomashuk, "Governance and regulation as an indispensable condition for developing the potential of rural areas," *Baltic Journal of Economic Studies*, vol. 5, no. 5, pp. 67–78, 2019.

[16] J. Jo, H. G. Kim, I. C. Park, B. C. Jung, and H. Yoo, "Modified Viterbi scoring for HMM-based speech recognition," *Intelligent Automation & Soft Computing*, vol. 25, no. 2, pp. 351–358, 2019.

[17] Z. Zeng, "Implementation of embedded technology-based English speech identification and translation system," *Computer Systems Science and Engineering*, vol. 35, no. 5, pp. 377–383, 2020.

[18] S. A. Banihashemi, M. Khalilzadeh, A. Shahraki, M. R. Malkhalifeh, and S. S. R. Ahmadizadeh, "Optimization of environmental impacts of construction projects: a time-cost-quality trade-off approach," *International Journal of Environmental Science and Technology*, vol. 18, no. 3, pp. 631–646, 2021.

[19] M. Ebrahimi, H. Nejadsoleymani, and M. R. Mansouri Daneshvar, "Land suitability map and ecological carrying capacity for the recognition of touristic zones in the Kalat region, Iran: a multi-criteria analysis based on AHP and GIS," *Asia-Pacific Journal of Regional Science*, vol. 3, no. 3, pp. 697–718, 2019.

[20] F. Asur, "An evaluation of visual landscape quality of coastal settlements: a case study of coastal areas in the Van lake basin (Turkey)," *Applied Ecology and Environmental Research*, vol. 17, no. 2, pp. 1849–1864, 2019.

[21] S. Yang, T. Gao, J. Wang, B. Deng, D. Landsdell, and B. Linares-Barranco, "Efficient spike-driven learning with dendritic event-based processing," *Frontiers in Neuroscience*, vol. 15, p. 97, 2021.

[22] C. Wang, C. Chen, Q. Pei, N. Lv, and H. Song, "Popularity incentive caching for vehicular named data networking," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2020.

[23] X. X. Liu, H. H. Song, and A. Liu, "Intelligent UAVs trajectory optimization from space-time for data collection in social networks," *IEEE Transactions on Network Science and Engineering*, p. 1, 2020.

[24] J. Yang, M. Xi, B. Jiang, J. Man, Q. Meng, and B. Li, "FADN: fully connected attitude detection network based on industrial video," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2011–2020, 2020.

[25] J. Yang, M. Xi, B. Jiang, and H. Song, "Robust six degrees of freedom estimation for IoT based on multibranch network," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2767–2775, 2021.