Hindawi
Mobile Information Systems
Volume 2022, Article ID 3541667, 9 pages
https://doi.org/10.1155/2022/3541667

Research Article

Machine Learning for Predictive Analytics in the Improvement of English Speech Feature Recognition

Yan Chen[1] **and Bukhari Martinuzzi**[2]

[1] *Business School, Chuzhou Polytechnic, Chuzhou 239000, Anhui, China*
[2] *International Atatürk-Alatoo University, Bishkek, Kyrgyzstan*

Correspondence should be addressed to Bukhari Martinuzzi; pro.bukhari@sm.cu.edu.kg

Received 2 July 2022; Revised 3 August 2022; Accepted 6 August 2022; Published 23 August 2022

Academic Editor: Praveen Kumar Reddy Maddikunta

The use of deep learning to improve English speaking has seen tremendous development in recent years. This study evaluates the noise that is present in the English speech environment, employs a two-way search method to select the optimum feature set, and applies a quick correlation filter to remove redundant features in order to increase the accuracy of English voice feature identification. In addition, this article designs a low-pass filter in the complex cepstrum domain to filter the room impulse response in order to obtain the estimated value of the complex cepstrum of the original speech signal. After doing so, the authors transform this estimated value into the time domain in order to obtain the estimated value of the original speech signal. In addition, this paper proposes a corresponding noise elimination model for the purpose of eliminating noise from English speech in a reverberant environment. It also designs a complex cepstrum domain filter in order to conduct simulation research on the different characteristics of the reverberation signal and the pure speech signal in the complex cepstrum domain. In conclusion, this study develops an English voice feature recognition model that is founded on a deep neural network. Furthermore, this paper uses experimental research to validate the validity of the algorithm model that was developed in this study.

1. Introduction

English speech enhancement based on the regression DNN network is proposed, and the experiment proves that the algorithm can achieve better performance than traditional English speech enhancement algorithms. However, although the English speech enhancement algorithm based on deep learning uses many noise types and training corpus in the training data preparation stage, there are still many problems in its promotion ability on real data, such as the distortion of English speech under low signal-to-noise ratio, the unstable effect of processing mismatched noise types, and mismatched speaking styles [1].

In the system environment disturbed by noise, the correct rate of English speech recognition is significantly reduced, resulting in the failure to achieve the ideal effect in practical applications, and the system is disturbed even more under the condition of low signal-to-noise ratio. In order to make the English speech signal detection system work normally, it is necessary to extract as much pure English speech as possible from the English speech signal contaminated by noise when the noise source is unknown. That is, under the premise of suppressing noise, the purpose of improving and protecting the quality of perceived English speech is achieved. This kind of English speech processing technology has great research significance and application value for the related fields of English speech signal processing. As far as the current English speech signal processing technology is concerned, the effect of English speech detection in a weak noise environment is relatively ideal. However, the detection performance drops sharply in a strong noisy environment. Therefore, the detection of English speech signals under the condition of low signal-to-noise ratio is still a subject to be studied in depth [2].

Analog signals are used to represent the English voice signal. However, because of the cut-off frequency, the

English voice is only present in the storage device as a digital signal as far as the English voice receiver is concerned. As a result, it starts by analysing the analogue English speech that has been digitally transformed, which typically entails amplification and gain control, prefiltering, sampling, quantization, and coding [3].

At present, English speech signal processing technology is developing rapidly in the field of information research, and its research scope involves cutting-edge scientific research projects, which has important research and application value. Moreover, informatization has become a basic requirement of modern society. In the civilian field, microphone array English speech signal processing technology is widely used in multimedia exhibition halls with large spaces and the hearing aid market. The English speech processing of the microphone array can adaptively control the beam direction, suppress interference signals in unknown directions in multiple directions, and have higher resolution. Therefore, in recent years, the development of adaptive processing technology has become more rapid, and the technology has also been used in other fields. However, the related algorithms of the English speech signal processing of the microphone array require a lot of floating-point operations. In current applications, most of them use DSP processors to perform operations on the collected signals. Although DSP has strong floating-point operations, it has disadvantages such as poor real-time serial operations and susceptibility to interference. Therefore, it is not competent for the more demanding processing system. This paper employs an FPGA-based English voice signal processing design to achieve this. The fact that the processor chip is inexpensive, compact, and capable of multichannel synchronous high-speed operation is a benefit. The development of FPGA-based English speech signal processing can thereby address the inadequacies of the current processing system and has significant implications for a wide range of applications.

In view of this, based on the deep neural network, this paper studies English speech feature recognition technology and proposes a reliable English speech feature recognition algorithm to provide a reference for subsequent English speech feature recognition.

2. Related Work

Research on endpoint detection and speech enhancement of noisy speech signals has been conducted for more than 50 years, and significant progress has been made during this period. Voice endpoint detection technology is proposed by [4], which is mainly applied to the time allocation of communication channels in the communication transmission system developed by it. The literature [5] proposed a system for reducing noise in the communication environment. The system introduces the concept that the input voice signal with noise is superimposed by the pure voice signal and the noise signal and divides the sample voice signal into multiple subbands for processing and analysis. The system is actually a spectral subtraction technique for now, but it is only implemented in the analog domain. Thanks to the rapid development of digital signal processing algorithms and DSP (digital signal processing) hardware, speech signal detection methods based on spectral improvements have been greatly developed, so speech signal noise reduction technology has made great progress. The literature [6] proposed a "spectrum shaping" method, which uses amplitude clipping in the filter bank of the speech signal preprocessing stage to remove low-level excitations. This low-level excitation is considered a noise signal. The literature [7] proposed spectral subtraction, which is implemented in the digital domain. Spectrum subtraction was applied to statistical spectrum estimate in [8]. Nearly and simultaneously, a technique that combines noise reduction and speech enhancement was suggested in [9]. The literature [10] proposed a voice endpoint recognition technique that establishes distinct thresholds to identify the starting point and ending point of the signal by combining the short-term energy of the speech signal with the short-term zero-crossing rate. The literature [11] explored endpoint detection performance in greater detail and developed algorithms for performance comparisons using several energy characteristics of the signal, including square energy, logarithmic energy, and absolute value energy. The optimum spectrum amplitude estimation and the best spectrum phase estimation are suggested by [12] using statistical prediction theory. The study's findings are frequently referenced in noise reduction studies, but, at the same time, the primary approach to noise reduction has changed to focus on the challenge of foreseeing the spectrum amplitude of pure speech signals. More statistical spectrum estimation approaches have been created by researchers, such as the minimum mean square error (MMSE) logarithmic spectrum amplitude estimation method, the maximum likelihood (ML) spectrum amplitude estimation method, and the maximum a posteriori (MAP) method. The Linear Predictive Coding (LPC) model and Kalman filter were utilised in [13] to reduce noise and raise the signal-to-noise ratio of speech signals. The literature [14] provided more endpoint detection algorithms through the frequency domain spectrum analysis of the voice signal after using the Fourier transform to get the frequency domain information of the voice signal. The literature [15] advocated for the speech signal's short-term stationarity and held that its parameter properties would be true over a brief period of time.

Segmentation methods based on LPC coefficients, methods based on speech parameters, and segmentation algorithms based on parameter filtering have been successively proposed. The literature [16] proposed an algorithm based on artificial neural network, through fast convergence to determine the different weights of the signal; its detection performance is significantly improved compared with the early statistical decision-making algorithm. Literature [17] proposes applying wavelet transform technology to speech signal detection, which greatly reduces the computational complexity of the algorithm.

The literature [18] researched the least square method. This blind system identification method uses the method of decomposing eigenvalues in the frequency band for processing. The literature [19] developed an adaptive filtering method. This method can combine Least Mean Square (LMS) and adaptive filtering methods. However, the disadvantage is that there are many restrictive conditions, the common zero point between channels will hinder this method, and the rank of the correlation matrix of the sound source signal is required to be maximized. The literature [20] studied the use of multichannel methods for linear prediction. This method is to diagonalize the covariance matrix of the speech signal to obtain the correlation characteristics of the signal. The literature [21] proposed using a virtual model to simulate the impulse response of the room. This method is based on the stability of the channel. However, under normal circumstances, the environment will change randomly, and it is difficult to meet this requirement, so this method is more difficult to implement.

3. English Speech Feature Recognition Algorithm Based on Deep Learning

This paper introduces the data set, data preprocessing, and extracted features, and two effective feature selection methods are used in feature selection. In addition, this paper uses three different classifiers and compares the classification effects.

We normalized all the data, as shown in the following formula:

$$\widetilde{a}(n) = \frac{a(n) - \mu(n)}{\sigma(n)}, \tag{1}$$

where $a(n)$ is the original sample, $\mu(n)$ and $\sigma(n)$ are the sample and standard deviation of the nth segment of data, each segment is 1 minute long, and $\widetilde{a}(n)$ is the normalized sample.

After preprocessing, each piece of data is equally segmented, and each segment is 1 minute long, and then features are extracted from each segment of the data. In this paper, 16 features are extracted from the single-channel ECG signal.

3.1. Time Domain Characteristics. The mean value of the RR interval without detrending, the mean value of the detrending RR interval, the standard deviation of the RR interval, the maximum value of the RR interval, the minimum value of the RR interval, and other features are extracted in this study based on the time domain. The fraction of RR intervals where the distance between two adjacent RR intervals is greater than 50 ms, the range of RR intervals, the root mean square of the distance between adjacent RR intervals, and the standard deviation of the distance between adjacent RR intervals are all factors to consider.

3.2. Frequency Domain Characteristics. In addition to the time domain, this paper also extracts a set of important frequency domain features. In order to extract the spectral characteristics of the RR signal, this paper performs fast Fourier transform (FFT) processing on the RR sequence and obtains four frequency domain characteristics: the power value of the extremely low frequency band, the power value of the low frequency band, and the power of the high frequency band.

3.3. Nonlinear Characteristics. In addition to time domain features and frequency domain features, this paper also extracts two nonlinear features: sample entropy and spectral entropy.

Multiscale entropy (MSE) is used to describe the structural complexity of time series. Many kinds of entropy can be used to calculate multiscale entropy, such as approximate entropy and fuzzy entropy under various time granularities. Multiscale entropy is increasingly used in sleep analysis. In this paper, sample entropy (SampEn) is used as the core of multiscale entropy calculation.

After the signal $\{x_i, i = 1\colon N\}$ of N data points is given, a coarse-grained time series $\{y(t)\}$ is first generated, where t is the scale factor. The ECG signal is divided into a nonoverlapping window of length t 1 : 1, and the average value is calculated.

$$y^{(t)} = \frac{1}{t}\sum_{i=(j-1)t+1}^{jt} x_i, \quad 1 \le j \le \frac{N}{t}, \tag{2}$$

Therefore, $y^{(1)}$ is the original signal, and $y^{(t)}$ is the coarse-grained sequence obtained by dividing the original sequence into windows of length t.

The calculation steps of sample entropy (SampEn) are as follows:

First, the coarse-grained time series form a set of m-dimensional vectors in order (m is the number of mode bits, and m is set to 2 in this paper): $x(i) = \{x(i), x(i+1),,, x(i+m-1)\}$, $(i = 1, 2, \ldots, N - m + 1)$.

We define the distance between $x(i)$ and $x(j)$ as $d[x(i), x(j)]$, which is the largest difference between the two elements; namely,

$$d[x(i), x(j)] = \max_{k=0\longrightarrow m-1}\{x(i+k) - x(j+k)\}. \tag{3}$$

For each value of i, we count the number n_i^m of $d[x(i), x(j)] < r$, $\begin{matrix} i = 1, 2, \ldots, N - m + 1, \\ j = 1, 2, \ldots, N - m + 1 \end{matrix}$ and $j \ne i$. Then, we calculate the ratio of it to the total number of distance $N - m$, denoted by

$$C_i^m(r) = \frac{n_i^m}{N - m}. \tag{4}$$

Then, the average value of $C_i^m(r)$ is

$$C^m(r) = \frac{1}{N - m + 1}\sum_{i=1}^{N-m+1} C_i^m(r). \tag{5}$$

The algorithm adds 1 to the dimension to become $m + 1$ and repeats the previous steps to count $C^{m+1}(r)$.

Finally, the calculation formula of sample entropy SampEn is

$$\mathrm{SampEn} = -\mathrm{In}\frac{C^{m+1}(r)}{C^m(r)}. \tag{6}$$

Spectral SpecEn describes the flatness of the power spectral density (PsD) and indirectly reflects the irregularity of the time series. Therefore, the larger the value of SpecEn, the flatter the shape of the PSD, and, accordingly, the more irregular it is distributed in the time domain. Conversely, the smaller the value of SpecEn, the denser the frequency spectrum and the lower the degree of irregularity of the PSD in the time domain distribution. It is also necessary to extract the spectral entropy as a feature.

In the sample training process, as the number of features increases, the length of time it takes to evaluate the features and train the model, as well as the model's complexity and promotion ability, all decreases. By removing unnecessary and duplicate features, feature selection can lower operating complexity.

This study divides the feature selection process into two phases. The optimum feature set for classification is first selected using the bidirectional search (BDS) algorithm, and the redundant features are then removed using the quick correlation filter.

Sequence forward selection (SFS) and sequence backward selection (SBS) are combined in the first step of the bidirectional search (BDS) method.

3.3.1. Bidirectional Search (BDS) Algorithm. Sequence forward selection (SFS) : add each feature to an empty set A one by one in turn. Each time a feature is added, the accuracy of the feature classification in A is calculated. If the accuracy is higher than before adding, the feature is valid and is kept in A; otherwise, the feature is invalid, and the feature is removed from A.

Sequence backward selection (SBS) : remove each feature one by one from the full set S and calculate the accuracy of the feature classification in *s* after removing a feature. If the accuracy is higher than before adding, continue; otherwise, keep the feature in S.

Bidirectional search (BDS) : use forward and backward sequence selection methods to search at the same time. When the results of the two process searches are the same feature subset, the search stops.

3.3.2. mRMR Algorithm. In the second stage, in order to evaluate the synergy between features and construct a set of optimal features, this paper adopts a filtering method based on mutual information and minimum redundancy and maximum correlation (mRMR) criteria.

The mRMR algorithm is based on mutual information. When two random variables x and Y are given and their probability density functions are $p(x)$, $p(y)$, and $p(x, y)$ respectively, the mutual information is

$$I(x; y) = \int\int p(x, y)\log\frac{p(x, y)}{p(x)p(y)}\,\mathrm{d}x\mathrm{d}y. \tag{7}$$

The goal of the algorithm is to find a feature subset containing $m(x_i)$ features.

The biggest correlation is

$$\max D(S, c), D = \frac{1}{|s|}\sum_{x_i\in s} I(x_i, c), \tag{8}$$

where x_if is the i-th feature, C is the categorical variable, and S is the feature subset.

The minimum redundancy is

$$\min R(S), R = \frac{1}{|s|^2}\sum_{x_i,x_j\in s} I\left(x_i, x_j\right). \tag{9}$$

Objective function addition integration:

$$\max\Phi(D, R), \Phi = D - R. \tag{10}$$

That is,

$$\max_{x_j\in X-S_{m-1}}\left[I\left(x_j, c\right) - \frac{1}{m-1}\sum_{x_i\in S_{m-1}} I\left(x_j, x_i\right)\right]. \tag{11}$$

Among them, X represents the complete set of feature x_j , s represents the set of selected feature x_i (size m), C represents the class, and I represents the mutual information. The definition of I is as follows:

$$I(x; y) = \int\int p(x, y)\log\frac{p(x, y)}{p(x)p(y)}\,\mathrm{d}x\mathrm{d}y. \tag{12}$$

Among them, $p(x)$, $p(y)$, and $p(x, y)$ are probability density functions. These three functions are estimated by a kernel density estimator based on adaptive diffusion.

This paper uses support vector machine (SVM), AdaBoost, and random forest three classifiers to classify English speech features.

AdaBoost Method. In addition to SVM, this paper also uses the AdaBoost (AB) method. Boosting algorithm has a good classification effect. Boosting is an iterative algorithm whose purpose is to combine several classification models and integrate them into one classification model. This integration method is based on the weighted voting of the same classifier.

AdaBoost (AB) is a widely used boosting algorithm, which was first proposed by Freund and Schapire. AB can be used with other classifiers, but if AB is applied to a complex classifier, the prediction performance of new data will be greatly affected; that is, the ability of promoting it will be lost. Therefore, when the weak classifier is applied to the AB algorithm, the effect will be better.

After every m iterations, the AB algorithm reassigns a new weight w_k^m for each feature vector x_k in the training set. Therefore, the m-th weak classifier will use the corresponding weights for training. Then, its classification performance is estimated with the error ε_m. This error is used to determine the weighted voting result of the m-th weak classifier.

Therefore, the smaller the error ε_m in these classifiers, the greater the contribution to the final classification. At the end of the iteration, the weight of the misclassified sample will be updated to w_k^{m+1}. Then, the weights of all samples will be standardized to maintain the original distribution.

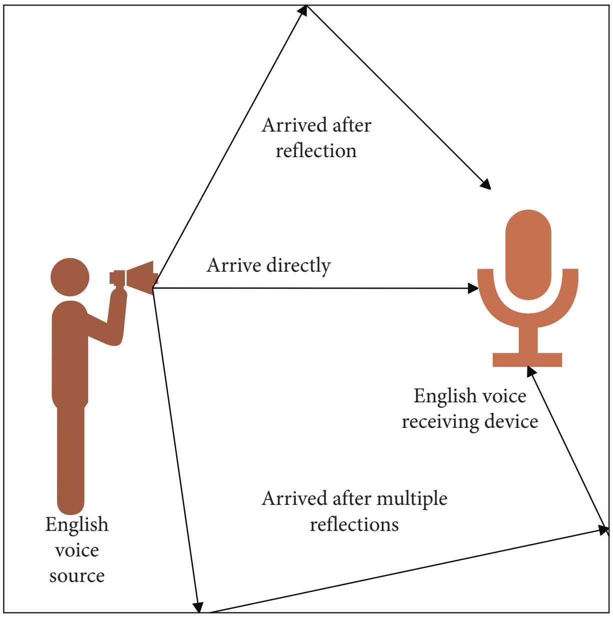

Figure 1: Schematic diagram of the English speech reverberation process in the classroom.

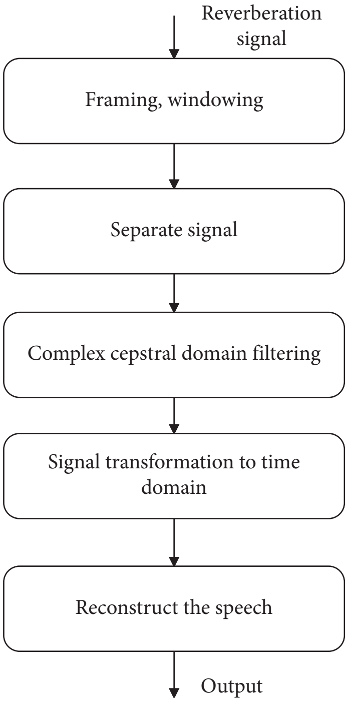

Figure 2: Flow chart of dereverberation.

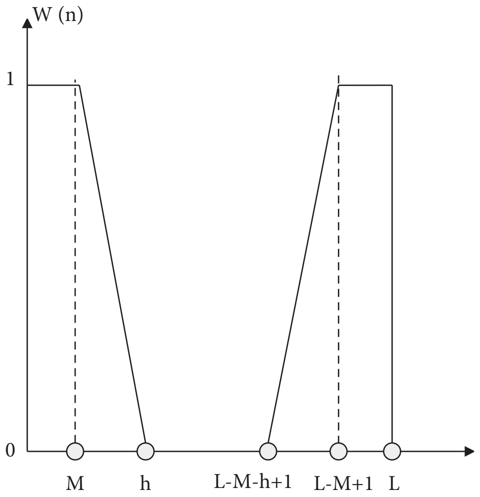

Figure 3: Schematic diagram of complex cepstral domain filter.

In this algorithm, the error ε_m of the m-th iteration is defined as the sum of the weights of the misclassified samples divided by the sum of the weights of all the samples in the current iteration.

$$\varepsilon_m = \frac{\sum_{k=1}^{N_{\text{trainning}}} w_k^m \,(\text{miss})}{\sum_{k=1}^{N_{\text{trainning}}} w_k^m}. \tag{13}$$

Random Forest. Random forest (RF) is a combination of multiple decision tree classifiers, each of which depends on an independently sampled random vector. Every decision tree in a random forest has the same distribution. As the number of decision trees in the random forest increases, the error of the random forest generated results gradually converges. The error of the random forest generated results depends on the strength of each independent decision tree in the forest and the relationship between the trees.

4. English Speech Feature Recognition System Based on Deep Neural Network

When performing English speech recognition in a classroom or in a relatively closed place, some of the sound waves emitted by the sound source are directly received by the microphone, and the other part will be reflected and absorbed after reaching the indoor walls, ceiling, ground, and other obstacles [22]. The attenuation of the sound signal after reflection is relatively small. Due to the different materials of various obstacles, the reflection coefficient is also different. In addition, the strength of the sound energy received by the obstacle is different, the signals received by the microphone will have a large amplitude compared with the original signal, and the phase will be different. From the reverberation process shown in Figure 1, it can be seen that reverberation is different from irrelevant external interference signals such as noise. The reverberation signal originates from the sound source signal and is a regular interference signal [23].

According to research on the complex cepstrum of the speech signal, the positions of the complex cepstrum of the sound source signal and the room's impulse response are different when the reverberant speech signal is translated into the complex cepstrum domain. While the latter is concentrated at both ends, the former is mostly concentrated closer to the midway point [24]. The estimated value of the complex cepstrum of the original speech signal must therefore be obtained by designing a low-pass filter in the

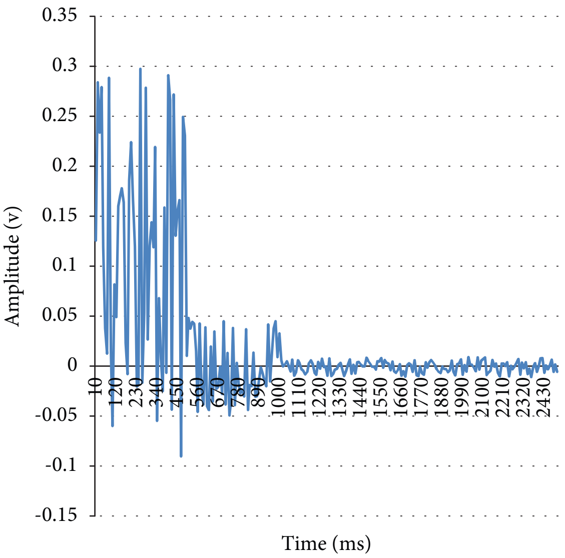

Figure 4: Impulse response function.

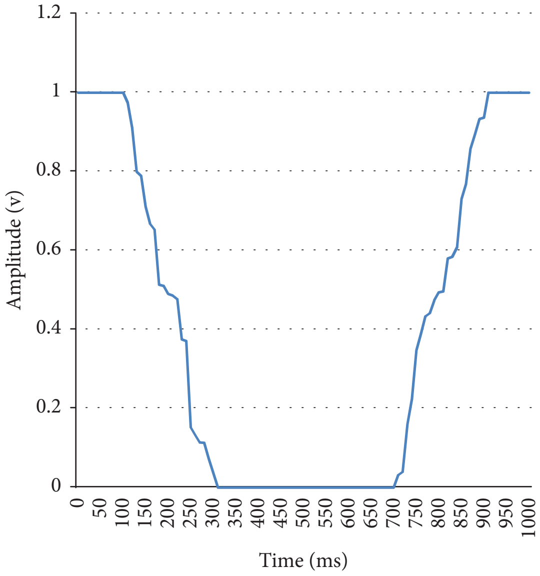

Figure 5: Low-pass filter.

complex cepstrum domain to filter the room impulse response, and this estimated value must then be transformed into the time domain to obtain the estimated value of the original speech signal. Figure 2 depicts the extensive cepstrum dereverberation procedure in this work.

Designing a complex cepstrum domain filter is an important part of the process of speech signal dereverberation. The complex cepstrum domain filter is a low-pass filter in a broad sense. Moreover, its parameters determine the performance of dereverberation, including three parts, namely, the pass band, the transition band, and the stop band. Figure 3 shows the filter schematic diagram.

Among them, L is the length of the filter, M is the cut-off point of the passband, h is the length of the transition band, and h(n) is the transition band function. When M is 1/16 of h and h is 1/8 of L, the best dereverberation evaluation index is obtained, and the dereverberation effect is the best.

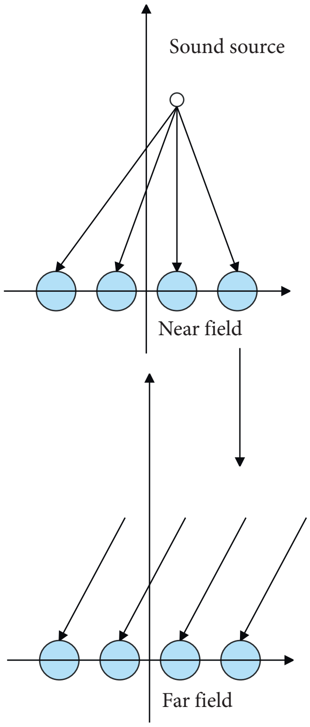

Figure 6: The near-field and far-field models of the microphone array.

This paper downloads an English voice from the officially recognized voice library. The sampling frequency is 44100 Hz, and the length, width, and height of the room used in the experiment are 5m, 4m, and 3m, respectively. Moreover, this paper uses the mirror image method to simulate the room impulse response, and the room impulse response function is shown in Figure 4. The collected voice is convolved with the simulated impulse response function to obtain the reverberant voice, and the reverberant voice is framed and then a Hamming window is added. Among them, the frame length is 1024, and the frame shift is 1/4 of the frame length.

As seen in Figure 5, this filter is a low-pass filter appropriate for the cepstrum domain. When the highest cut-off point for the filter is 1/256 of the frame length and the bandwidth of the transition band is 1/16 of the frame length, it is discovered that good evaluation results for the speech signal obtained after dereverberation may be obtained.

According to the distance from the sound source to the microphone array, it is divided into a near-field model and a far-field model of the microphone array. When the signal source is far from the array, the wave path difference of the signal reaching each element is relatively small, and the signal can be treated as a plane wave model. The difference is that when the signal source is close to the microphone array, the signal reaches the array element in the microphone array with a larger amplitude difference. At this time, the waveform arriving at the array should be a spherical wave model. Figure 6 shows the near-field and far-field models of the microphone array.

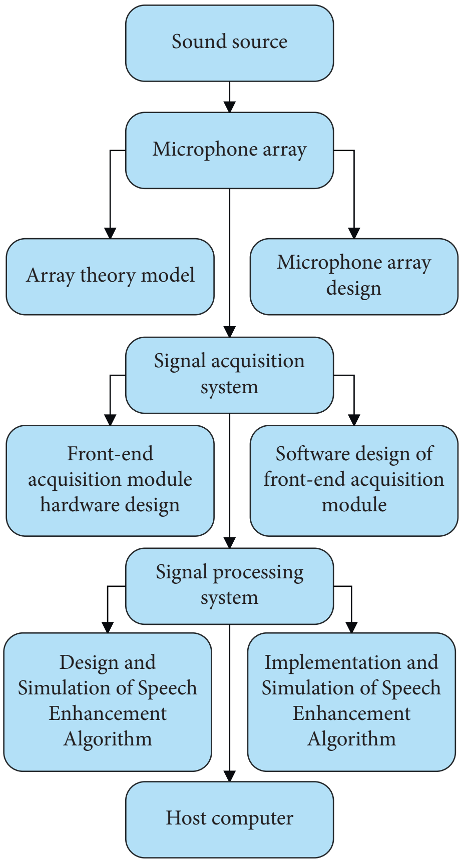
Sound source
Microphone array
Array theory model
Microphone array design
Signal acquisition system
Front-end acquisition module hardware design
Software design of front-end acquisition module
Signal processing system
Design and Simulation of Speech Enhancement Algorithm
Implementation and Simulation of Speech Enhancement Algorithm
Host computer

Figure 7: Block diagram of the overall implementation scheme of the English speech recognition system.

The overall implementation scheme of the FPGA-based microphone array signal processing system is shown in Figure 7. First, a microphone array is designed as the voice signal collection terminal. This paper uses 4 low-cost omnidirectional electret microphones as the elements of the microphone array to convert the voice signal into an analog signal output. Then, a signal acquisition system with signal acquisition and AD conversion functions is designed.

The model in this paper is based on the foundation of deep neural network. The results of the deep neural network in this paper are shown in Figure 8.

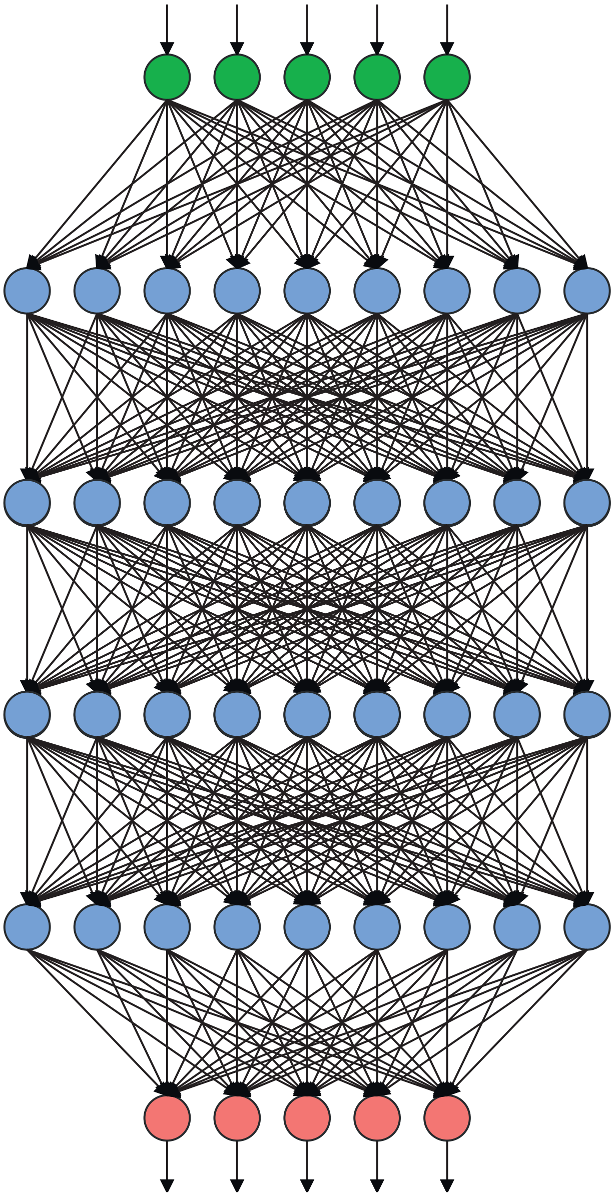

Figure 8: Deep neural network model.

5. Performance Verification of English Speech Feature Recognition Model Based on Deep Neural Network

This study uses deep neural networks to construct a model for English speech feature recognition. This model can perform English voice denoising using a neural network approach in order to accomplish the recognition of English speech features even in situations when there is classroom reverberation. As a result, this work initially assesses the impact of English speech denoising before counting the impact of English speech feature recognition in the system performance test. In order to determine the denoising effect of English speech, this study collects numerous sets of English speech data via the network and runs tests with the system that it has built, as shown in Table 1 and Figure 9.

From the analysis results of the above chart, it can be seen that the English speech feature recognition model based on the deep neural network constructed in this paper has a better effect. After that, this paper conducts the evaluation of the English speech feature recognition effect of the system constructed in this paper. The results obtained are shown in Table 2 and Figure 10.

Table 1: Statistical table of the accuracy of English speech denoising.

Num	Speech denoising (%)	Num	Speech denoising (%)	Num	Speech denoising (%)
1	92.36	27	90.05	53	90.55
2	92.08	28	92.73	54	91.93
3	92.30	29	92.10	55	93.56
4	91.32	30	91.70	56	91.55
5	93.20	31	91.08	57	92.61
6	93.01	32	93.01	58	93.78
7	91.71	33	90.74	59	93.75
8	91.04	34	90.07	60	90.52
9	90.81	35	93.46	61	92.77
10	93.95	36	92.80	62	93.81
11	90.32	37	90.36	63	93.08
12	92.38	38	92.63	64	91.37
13	92.67	39	90.39	65	92.01
14	94.00	40	92.76	66	90.85
15	91.18	41	91.22	67	93.56
16	92.38	42	91.31	68	90.84
17	91.82	43	93.86	69	92.42
18	93.06	44	91.47	70	93.07
19	92.95	45	91.02	71	92.44
20	91.45	46	90.59	72	91.88
21	92.35	47	93.37	73	92.97
22	90.92	48	93.37	74	91.33
23	92.43	49	94.00	75	93.59
24	92.19	50	91.47	76	91.70
25	91.50	51	90.64	77	91.22
26	93.42	52	93.88	78	90.24

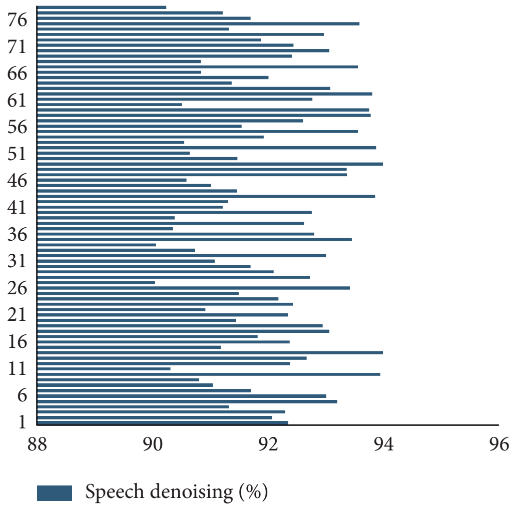

Figure 9: Statistical diagram of the accuracy of English speech denoising.

Table 2: Statistical table of the speech feature recognition effect of the English speech feature recognition system.

Num	Speech recognition (%)	Num	Speech recognition (%)	Num	Speech recognition (%)
1	90.81	27	90.86	53	90.19
2	90.22	28	89.98	54	88.06
3	85.37	29	90.39	55	87.00
4	89.84	30	87.43	56	85.79
5	87.59	31	88.71	57	85.76
6	88.45	32	88.23	58	84.77
7	84.96	33	87.94	59	88.61
8	89.07	34	84.20	60	90.07
9	90.20	35	88.74	61	89.25
10	87.86	36	87.21	62	89.12
11	87.34	37	86.26	63	89.99
12	89.82	38	87.61	64	87.54
13	84.29	39	90.37	65	90.44
14	85.20	40	84.74	66	88.88
15	88.38	41	84.42	67	89.52
16	84.80	42	88.40	68	84.61
17	87.95	43	87.67	69	88.45
18	91.00	44	87.02	70	87.84
19	88.26	45	88.66	71	89.05
20	87.04	46	88.33	72	89.48
21	90.30	47	90.07	73	88.31
22	86.67	48	87.18	74	85.96
23	89.35	49	89.69	75	89.78
24	89.73	50	84.79	76	86.36
25	86.64	51	90.79	77	86.84
26	88.28	52	90.83	78	85.82

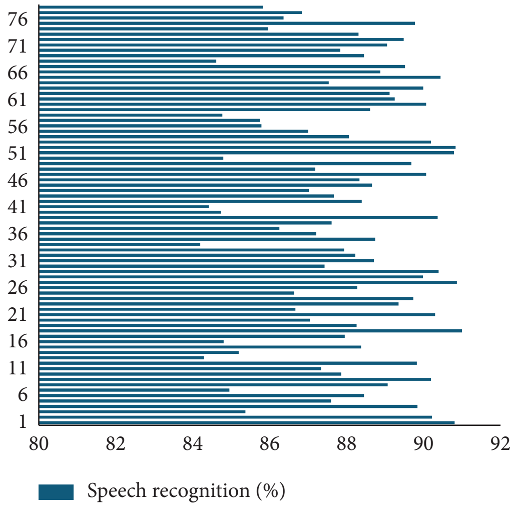

Figure 10: Statistical diagram of the speech feature recognition effect of the English speech feature recognition system.

From the above experimental research results, it can be seen that the English speech feature recognition system constructed in this paper has a certain effect.

6. Conclusion

This paper studies the English speech detection algorithm based on the nonstationary strong noise environment. The windowing of the English speech signal can make the speech signal processing easier, and different window functions have different effects. Linear predictive analysis includes autocorrelation method and covariance method. The covariance approach is less reliable than the autocorrelation method, which is better suited for interpreting English voice signals. In this study, the filter bank addition and overlap addition are introduced for the short-term synthesis of English voice signals. Additionally, the concatenation and addition approach is chosen to handle the voice signal due to its simplicity after evaluating the two methods' degree of complexity. This work also conducts simulation research on the various properties of the reverberation signal and pure speech signal in the complex cepstrum domain, examines the basic idea of complex cepstrum domain filtering, and builds a complex cepstrum domain filter. Finally, this paper constructs an English speech feature recognition model based on deep neural network and verifies the reliability of the algorithm model through experimental research [25, 26].

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] P. H. Kumar and M. N. Mohanty, "Efficient feature extraction for fear state analysis from human voice," *Indian Journal of Science & Technology*, vol. 9, no. 38, pp. 1–11, 2016.

[2] R. Rhodes, "Aging effects on voice features used in forensic speaker comparison," *International Journal of Speech Language and the Law*, vol. 24, no. 2, pp. 177–199, 2017.

[3] Q. K. D. Ngoc and D. Hien Thanh, "A review of audio features and statistical models exploited for voice pattern design," *Computer Science*, vol. 03, no. 2, pp. 36–39, 2015.

[4] M. Sarria-Paja, M. Senoussaoui, and T. H. Falk, "The effects of whispered speech on state-of-the-art voice based biometrics systems," in *Proceedings of the 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering*, pp. 1254–1259, IEEE, Halifax, NS, Canada, May 2015.

[5] A. Leeman, H. Mixdorff, M. O'Reilly, M. J. Kolly, and V. Dellwo, "Speaker-individuality in Fujisaki model f0 features: implications for forensic voice comparison," *International Journal of Speech Language and the Law*, vol. 21, no. 2, pp. 343–370, 2015.

[6] A. K. Hill, R. A. Cárdenas, J. R. Wheatley et al., "Are there vocal cues to human developmental stability? Relationships between facial fluctuating asymmetry and voice attractiveness," *Evolution and Human Behavior*, vol. 38, no. 2, pp. 249–258, 2017.

[7] M. Woźniak and D. Polap, "Voice recognition through the use of Gabor transform and heuristic algorithm," *International Journal of Electronics and Telecommunications*, vol. 63, no. 2, pp. 159–164, 2017.

[8] T. Haderlein, M. Döllinger, V. Matousek, and E. Noth, "Objective voice and speech analysis of persons with chronic hoarseness by prosodic analysis of speech samples," *Logopedics Phoniatrics Vocology*, vol. 41, no. 3, pp. 106–116, 2015.

[9] S. S. Nidhyananthan, K. Muthugeetha, and V. Vallimayil, "Human recognition using voice print in LabVIEW," *International Journal of Applied Engineering Research*, vol. 13, no. 10, pp. 8126–8130, 2018.

[10] F. L. Malallah, K. N. Y. M. G. Saeed, S. D. Abdulameer, and A. W. Altuhafi, "Vision-based control by hand-directional gestures converting to voice," *International Journal of Scientific & Technology Research*, vol. 7, no. 7, pp. 185–190, 2018.

[11] M. Sleeper, "Contact effects on voice-onset time in Patagonian Welsh," *Journal of the Acoustical Society of America*, vol. 140, no. 4, p. 3111, 2016.

[12] G. Mohan, K. Hamilton, A. Grasberger, A. C. Lammert, and J. Waterman, "Realtime voice activity and pitch modulation for laryngectomy transducers using head and facial gestures," *Journal of the Acoustical Society of America*, vol. 137, no. 4, p. 2302, 2015.

[13] C. T. Herbst, S. Hertegard, D. Zangger-Borch, and L. Per-Åke, "Freddie Mercury—acoustic analysis of speaking fundamental frequency, vibrato, and subharmonics," *Logopedics Phoniatrics Vocology*, vol. 42, no. 1, pp. 1–10, 2016.

[14] J. Al-Tamimi, "Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: implications for formal representations," *Laboratory Phonology*, vol. 8, no. 1, pp. 1–40, 2017.

[15] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition IEEE/ACM Transactions on audio," *Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[16] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315–1329, 2016.

[17] M. A. Carrier, "Product hopping," *Journal of Commercial Biotechnology*, vol. 23, pp. 52–60, 2017.

[18] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[19] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[20] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[21] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: a survey," *Speech Communication*, vol. 56, no. 3, pp. 85–100, 2014.

[22] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.

[23] Y. miao, Y. Huang, and Z. Da, *English Speech Feature Recognition Based on Digital Means*, Research Square, Durham, NC USA, 2021.

[24] P. J. Pitts, "21st century pharmacovigilance: intuition, science, and the role of artificial intelligence," *Journal of Commercial Biotechnology*, vol. 23, pp. 3–6, 2017.