Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 4421520, 10 pages
https://doi.org/10.1155/2022/4421520

Research Article

English Speech Feature Recognition-Based Fuzzy Algorithm and Artificial Intelligent

Yuji Miao, Haiying Liu, and Shan Gu

Faculty of International Languages, Qinggong College, North China University of Science and Technology, Tangshan, Hebei 064000, China

Correspondence should be addressed to Haiying Liu; 171849022@masu.edu.cn

Received 7 February 2022; Revised 29 March 2022; Accepted 15 April 2022; Published 4 May 2022

Academic Editor: Rashid A Saeed

It is necessary to study the application of digital technology in English speech feature recognition. This paper combines the actual needs of English speech feature recognition to improve the digital algorithm. Moreover, this paper combines fuzzy algorithm to analyze English speech features, analyzes the shortcomings of traditional algorithms, proposes the fuzzy digitized English speech recognition algorithm, and builds an English speech feature recognition model on this basis. In addition, this paper conducts time-frequency analysis on chaotic signals and speech signals, eliminates noise in English speech features, improves the recognition effect of English speech features, and builds an English speech feature recognition system based on digital means. Finally, this paper conducts grouping experiments by inputting students' English pronunciation forms and counts the results of the experiments to test the performance of the system. The research results show that the method proposed in this paper has a certain effect.

1. Introduction

English speech feature recognition plays an important role in supporting English learning. From a practical point of view, improving English speech feature recognition through digital means can effectively improve the effect of English learning. [1]. Language has lost its use as the most fundamental and natural form of communication. The keyboard confines human hands, which have been emancipated from walking upright. This is plainly unacceptable to human beings who have worked tirelessly to build a magnificent civilization [2]. As a result, practically at the same time as people use computers, humans attempt to connect with computers by speech in order for computers to comprehend human languages, but this has always shown to be ineffective compared to keyboards and mouse. Humanity has entered the post-PC age with the arrival of the multimedia and network era. A broad range of intelligent equipment is extensively employed in human production and existence at present time. Voice is once again a vital demand for human-machine communication due to its natural, rapid, stable, and dependable means of contact between people and intelligent terminals. The speech signal, which serves as the carrier of voice recognition, has also been a focus of study [3].

Moreover, it is a comprehensive subject formed on the basis of speech linguistics and digital signal processing, and it is closely related to disciplines such as psychology, physiology, computer science, communication and information science, pattern recognition, and artificial intelligence. At the same time, the research on speech signal processing has always been an important driving force for the development of digital signal processing technology. Many new methods of processing are first achieved in speech processing and then extended to other fields. For example, the birth and development of many high-speed signal processors are inseparable from the research and development of speech signal processing. The so-called speech signal processing is the use of digital signal processing technology to analyze and process the speech signal, which includes speech communication, synthesis, recognition, and speech enhancement. One of its purposes is to obtain some speech parameters reflecting the important characteristics of the speech signal through processing so as to transmit or store the speech signal information with high effect. The second purposes is to process certain calculations to meet the requirements of a certain purpose, such as artificially synthesizing speech, identifying the speaker, and identifying the content of the speech. These two purposes represent the two aspects of the theory and research of speech signal processing that are closely integrated. In addition, voice signal processing is one of the core technologies used in emerging fields such as information superhighway, multimedia technology, office automation, modern communications, and intelligent systems.

This paper uses digital means to study English speech feature recognition algorithms and analyzes actual cases to improve the effect of English speech feature recognition.

2. Related Work

The speech signal is a typical nonstationary random signal. Due to the wide variety of voice environments and the complexity of the voice signal itself, people will inevitably be interfered by noise introduced from the surrounding environment and transmission media, electrical noise inside the communication device, and even other speakers during voice communication. These interferences will eventually make the voice received by the receiver no longer pure original voice, but noisy voice polluted by noise. This causes the deterioration of the performance of the speech processing system, affects the recognition rate, and even causes the system to be completely unable to process speech. Moreover, these interference signals cause great interference to the effective information carried by the voice signal in the voice communication. As a result, the denoising processing of the speech signal is produced [4].

The literature [5] looked into the linked issues of a watermark's anti-interference capacity in digital-to-analog conversion and utilized the ICA technique to differentiate two distinct sounds from an a cappella recording. Because of the excellent separation effect of the ICA method, the literature [6] was able to effectively achieve the interactive alteration of music recordings. The extraction of tiny target signals against a clutter background and the study of system security performance are all described in the literature [7]. The literature [8] conducted a study on air traffic control safety communication concerns using the pilot's voice and the aircraft's identify and produced excellent research findings. The literature [9] embedded watermark information in the telephone speech and confirmed the viability of the approach through simulated studies in order to develop the JP tele-phone network intrusion detection system. A recursive method based on neural simulation structure was presented in the literature [10]. The literature [11] provided a rather full study foundation for the issue of blind source separation, and it presented a joint diagonalization approach by examining the separability and uncertainty in the blind source separation algorithm. The literature-proposed BP neural network algorithm [12] has become one of the most extensively utilized neural network models. On the basis of information theory, the literature [13] devised an objective function based on information maximization and built a unified framework for the ICA method.

The literature [14] proposed a disguised voice hidden telephone system. Moreover, due to the consideration of the real-time nature of the system, it combines the classic hiding method to successfully realize the secure transmission of real-time voice. The literature [15] used audio as a hidden carrier to mask the secret voice signal, so as to enhance the signal's anti-low-pass filtering ability and anti-interference ability. The literature [16] derived an adaptive blind source separation switching algorithm based on kurtosis and used this algorithm to achieve blind separation of speech signals. The simulation experiment verifies that the algorithm has good separation performance. The literature [17] established a dual-channel blind source separation model on the basis of the 4th order cumulant and used the feature matrix joint approximate diagonalization algorithm to further realize the separation of the frequency hopping signal and the interference. Aiming at the problem of noise interference in multiple received signals, the literature [18] clearly pointed out that the signal can be denoised by wavelet first and then blindly separated. The literature [19] extensively detailed the blind source separation algorithm and further discussed the blind source separation algorithm's future growth path. The feature matrix joint approximation diagonalization approach was utilized in the literature [20] to blind source separation of the signal. At the receiving end, earlier blind source separation anti-interference approaches need a high number of antennas. The literature [21] presented a semi-blind separation anti-interference algorithm for direct spread communication based on the periodicity of the spreading code under a single channel as a solution to this issue.

3. Digital English Speech Processing Algorithm

3.1. Data Fuzzification. Fuzzification is the process of transforming numerical attributes into fuzzy attributes. Nowadays, many methods have been proposed to realize this transformation.

Generally speaking, it can be studied by experts in the field of specific discretization algorithm or by any fuzzy clustering algorithm. Clustering is a popular data mining technology, which allocates data instances to several groups called clusters, so that the instances belonged to different clusters are different as much as possible. The similarity between clusters and instances can be measured by distance, connectivity, and strength. Nonfuzzy clustering algorithms, such as DBSCAN or K-means, assign each instance to a cluster.

In the case of fuzzy clustering, each instance can be associated with more clusters with membership degree.

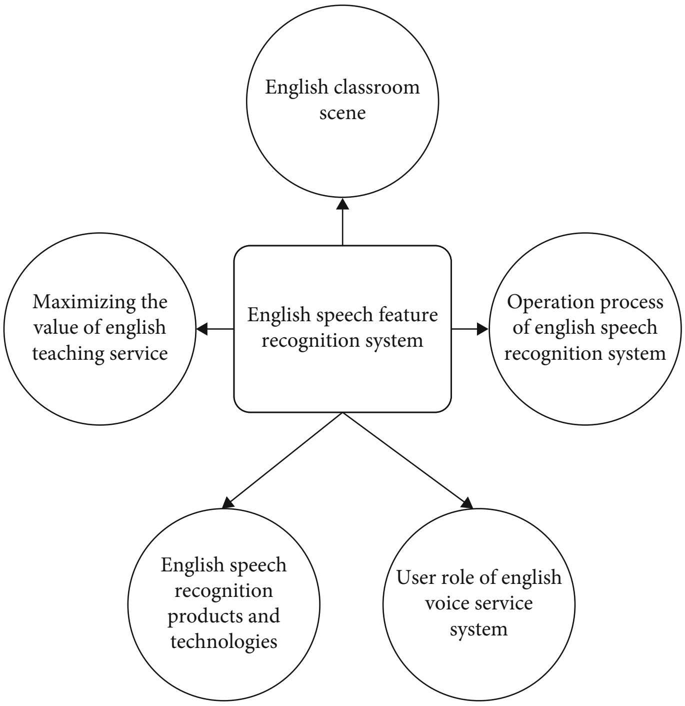
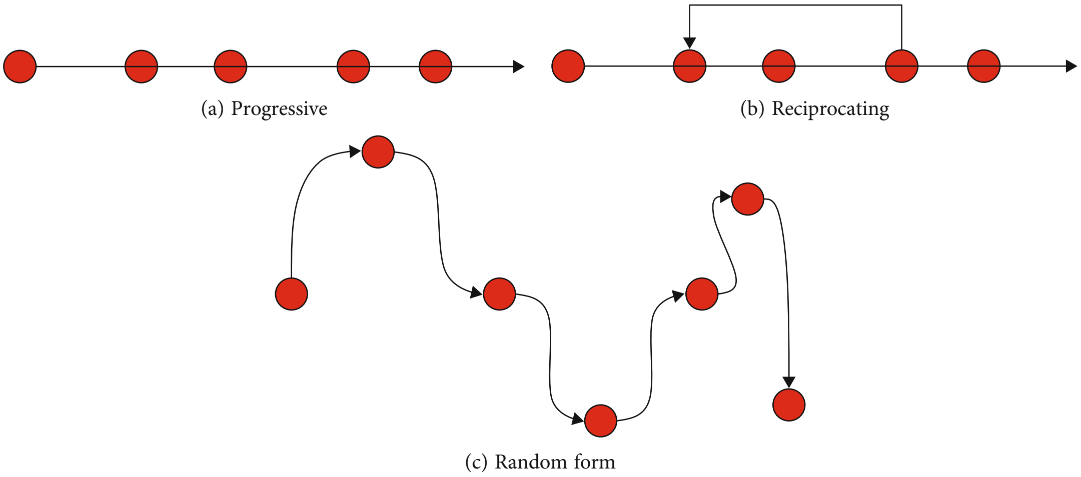

English classroom scene
Maximizing the value of english teaching service
English speech feature recognition system
Operation process of english speech recognition system
English speech recognition products and technologies
User role of english voice service system

Figure 1: Elements of the system.


(a) Progressive

(b) Reciprocating

(c) Random form

Figure 2: Interactive behavior path.

Membership indicates the degree to which the instance belongs to the cluster. One of the most commonly used fuzzy clustering algorithms is fuzzy C-means (FCM).

The purpose of this algorithm is to minimize the following criteria [22]:

$$\text{Minimize} \sum_{j=1}^{m_i} \sum_{i=1}^{k} \left(u_{i,j}\right)^2 d\left(a_i, c_j\right)^2. \quad (1)$$

Among them, $u_{i,j}$ represents the membership degree of the i-th instance to the j-th cluster, f is the ambiguity, and $d(a_i, c_j)$ is the distance metric.

$$d\left(a_i, c_j\right) = \sqrt{\left(a_i - c_j\right)^2} = \left|a_i - c_j\right|. \quad (2)$$

This algorithm defines the constraint of the membership degree $u_{i,j}$ of each level, and the sum of the membership degree of each cluster of a_i must be equal to 1. This constraint starts directly from the use of fuzzy C-means, but it

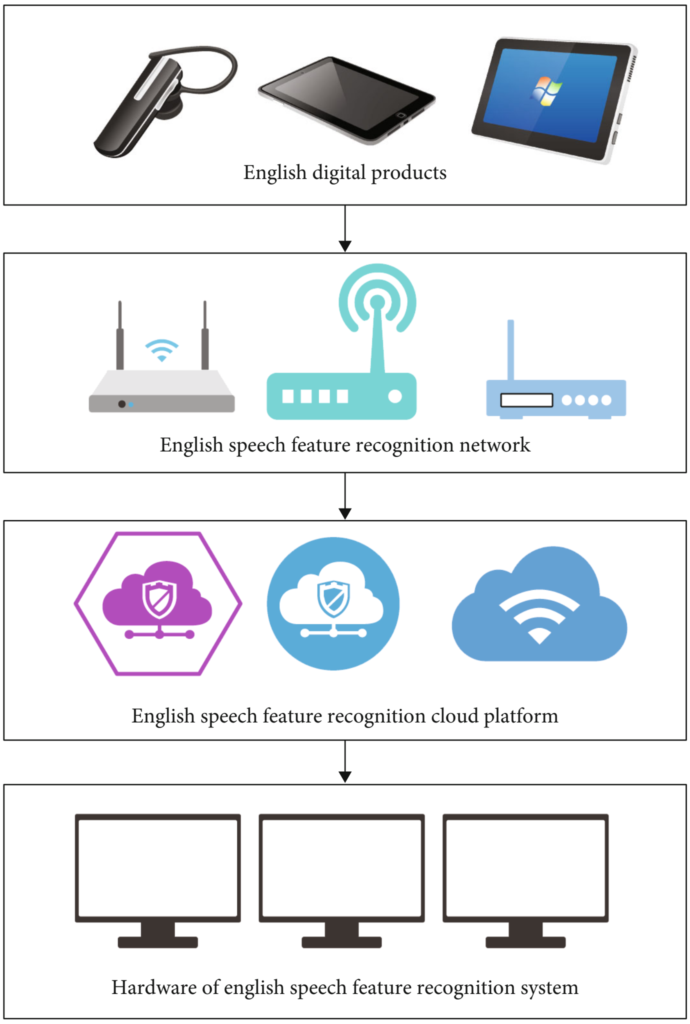

English digital products
English speech feature recognition network
English speech feature recognition cloud platform
Hardware of english speech feature recognition system

Figure 3: The application of the Internet of Things system in English speech feature recognition.

is also necessary to achieve FDT induction. The constraint can be expressed as the following formula:

$$\sum_{j=1}^{m_i} u_{i,j} = 1, \quad j = 1, 2, \cdots, k,$$
$$0 \le u_{i,j} \le 1, \quad i = 1, 2, \cdots, k\,; j = 1, 2, \cdots, m_i, \qquad (3)$$
$$0 \le \sum_{j=1}^{m_i} u_{i,j} \le k, \quad i = 1, 2, \cdots, k.$$

FCM assigns a membership degree $u_{i,j}$ to each cluster's instance of each cluster. The degree of membership is calculated based on the distance between the instance and the cluster. The membership degree of the i-th instance to the j-th cluster is as follows:

$$u_{i,j} = \frac{1/\left(d\left(a_i, c_j\right)\right)}{\sum_{t=1}^{m_i} 1/(d(a_i, c_t))}. \qquad (4)$$

When all the values of $a_i \in a$ are assigned to each cluster, the new center can be calculated by the following formula:

$$c_j = \frac{\sum_{j=2}^{k} \left(u_{i,j}\right)^2 a_i}{\sum_{j=1}^{k} \left(u_{i,j}\right)^2}. \qquad (5)$$

The scalar value can be converted into a degree of membership, and the continuous value can be converted into a fuzzy attribute mapping.

3.2. Fuzzy Decision Tree. DTS is a data mining tool for classification and prediction. The main purpose of classification is to map a new unknown instance to one of the predefined classes.

At present, there are many known DT sensing algorithms, such as ID3, C4.5, CHAID, cart, or CMIE-based fuzzy decision tree. ID3 has good performance in linguistic attributes, but it cannot deal with numerical attributes. The next problem with ID3 is that it tends to prefer attributes

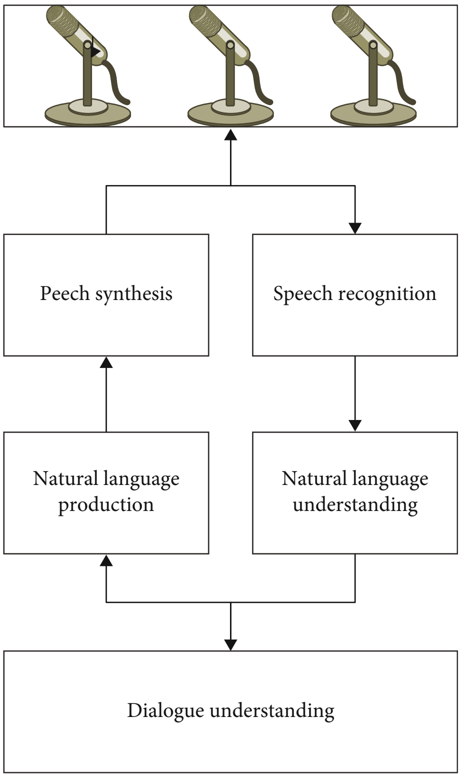


FIGURE 4: Intelligent voice-related technologies.

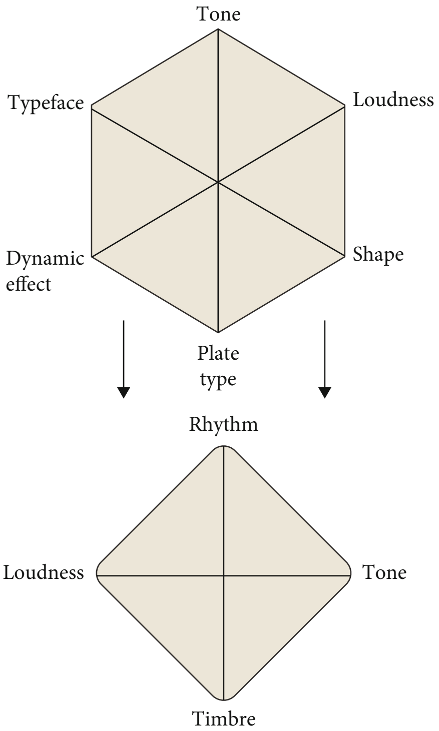


FIGURE 5: The difference between graphic interaction design and voice interaction design elements.

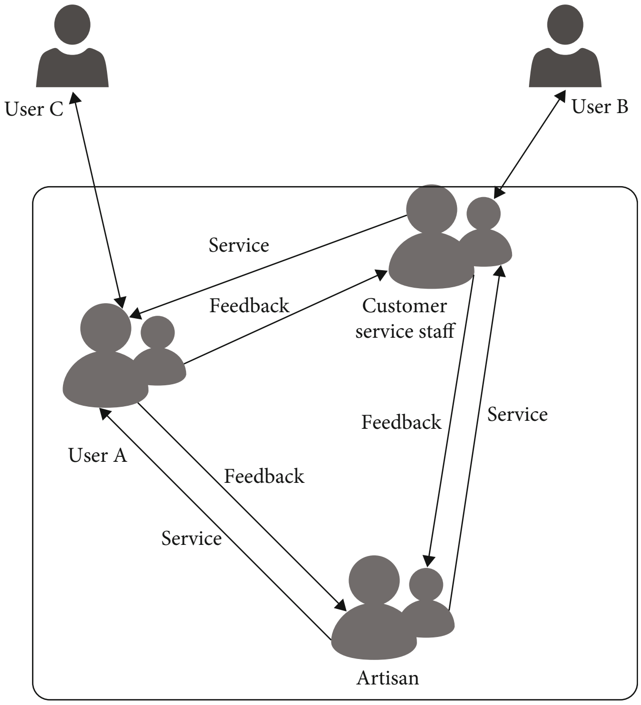


FIGURE 6: Map of key figures.

with more values. These problems are improved in the C4.5 algorithm. C4.5 deals with numerical attributes by defining split points, which divide numerical data into intervals. Unfortunately, it can have a negative impact on the performance of classification, especially when some values are close to the margin. In addition, the identification of interval boundary may not be completely correct. The introduction of fuzzy logic can improve this aspect. At present, various algorithms of fuzzy decision tree have been introduced. Many of them are based on traditional algorithms, such as ID3 to fuzzy correction. Other FDT guidance algorithms are based on CMIE proposed and applied in.

The associated attribute of each internal node selected by CMIE is defined as $I(B\,;A_{i_1j_1},A_{i_2j_2},\cdots,A_{i_{q-1}j_{q-1}},A_{i_qj_q})$. According to $U_{q-1}=A_{i_1j_1},A_{i_2j_2},\cdots,A_{i_{q-1}j_{q-1}},A_{i_qj_q}$, the sequence of attribute values is defined from the root of the tree to the node at the q-th level. The sequence U_{q-1} defines the path from the root to the node under investigation. The attribute of the maximum value of CMIE is associated with the investigated node in the obtained selection criterion, and the CMIE is divided by the entropy of the attribute, thereby preventing a larger number of attribute preferences. The standard has the following form:

$$\arg\max\left(\frac{I\left(B\,;A_{i_1j_1},A_{i_2j_2},\cdots,A_{i_{q-1}j_{q-1}},A_{i_qj_q}\right)}{H\left(A_{i_q}\right)}\right). \quad (6)$$

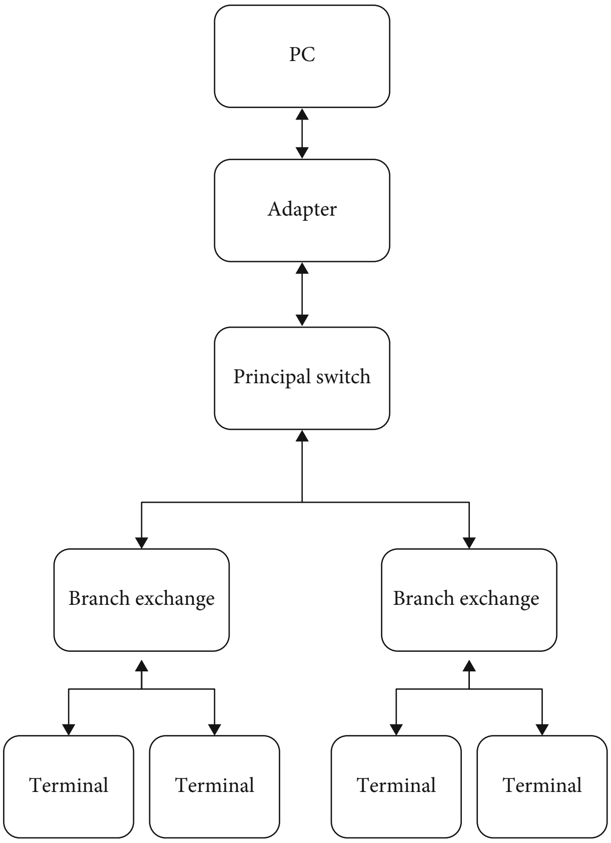


Figure 7: Block diagram of English speech feature recognition system.

Table 1: Statistical table of the effect of digital conversion of English speech.

No.	Digital effect	No.	Digital effect	No.	Digital effect
1	79.5	28	87.6	55	91.5
2	81.4	29	85.1	56	79.6
3	86.8	30	89.7	57	84.3
4	80.2	31	80.8	58	81.4
5	87.0	32	81.0	59	85.9
6	90.0	33	80.7	60	79.5
7	81.0	34	79.6	61	91.6
8	81.4	35	86.3	62	89.6
9	91.0	36	82.7	63	85.8
10	87.3	37	88.5	64	88.3
11	79.3	38	84.0	65	85.4
12	84.4	39	83.0	66	79.7
13	80.6	40	91.9	67	80.9
14	87.1	41	89.1	68	90.8
15	86.2	42	86.9	69	84.6
16	91.0	43	89.8	70	88.5
17	87.8	44	91.3	71	83.4
18	82.0	45	83.4	72	79.5
19	84.4	46	86.4	73	91.0
20	81.6	47	91.1	74	83.3
21	84.3	48	91.1	75	79.6
22	89.5	49	84.5	76	85.3
23	86.6	50	82.3	77	88.9
24	88.5	51	86.2	78	79.5
25	83.2	52	80.9	79	83.0
26	81.7	53	79.3	80	89.0
27	83.1	54	86.5		

The entropy of attribute A_{i_q} has the following definition:

$$H\left(A_{i_q}\right) = \sum_{=1}^{m_i} M\left(A_{i_q j_q}\right) * \left(\log_2 k - \log_2 M\left(A_{i_q j_q}\right)\right). \tag{7}$$

Overfitting is a problem for DT. Overfitting happens when the classifier overfits the data it detects. As a result, when the new data is unknown, DT's classification performance will be poorer. Overfitting in the decision tree refers to the tree that is induced fitting all of the cases in the data set. In this situation, it correctly identifies training examples, but it often incorrectly classifies fresh data. As a result, the pruning procedure is used, which involves replacing the tree's subtrees with leaves. The computational complexity of categorization is reduced through tree pruning. Leaves is divided into two categories, A and B, and used the pruning of the leaves to decide the leaf nodes during tree derivation. The minimal frequency of occurrence in a particular branch is represented by threshold A. The proportion of instances covered by the route to which the analysis node belongs is represented by frequency. When the following conditions are true, the node is converted to a leaf.

$$\alpha \geq \frac{M\left(A_{i_1 j_1} \times \cdots \times A_{i_q j_q}\right)}{k}. \tag{8}$$

Each node contains a confidence level that reflects the credibility of the output class. If the confidence is greater than the threshold parameter B, the node is transformed on the leaf. The mathematical form of this pruning criterion is as follows:

$$\beta \leq 2^{-I\left(B_j \middle| A_{i_1 j_1},\cdots,A_{i_z j_z}\right)}, \quad j = 1, \cdots, m_b, \tag{9}$$

$I(B_j|A_{i_1 j_1},\cdots,A_{i_z j_z})$ can be calculated by the following formula:

$$I\left(A_{i_1 j_1} \middle| A_{i_z j_z}\right) = H\left(A_{i_q}\right) = \log_2 M\left(A_{i_1 j_1}\right) - \log_2 M\left(A_{i_z j_z} \times A_{i_1 j_1}\right). \tag{10}$$

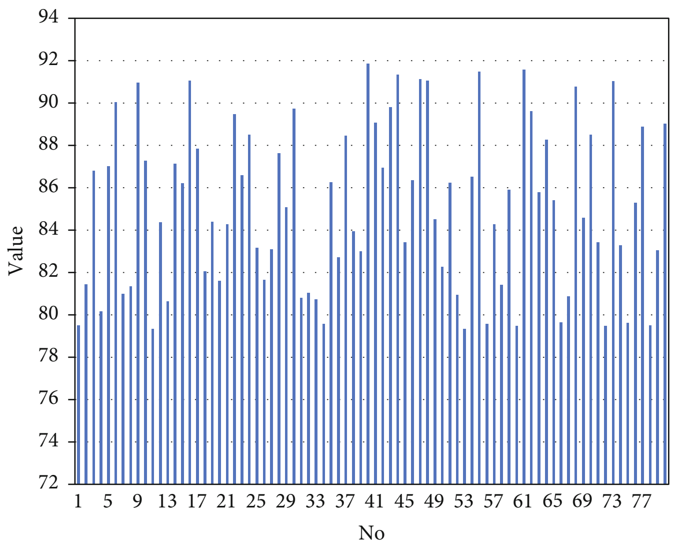


Figure 8: Statistical diagram of the effect of digital conversion of English speech.

Table 2: Statistical table of the accuracy of English feature recognition.

No.	Feature recognition effect	No.	Feature recognition effect	No.	Feature recognition effect
1	83.5	28	90.7	55	88.0
2	85.5	29	77.1	56	93.2
3	93.8	30	81.5	57	91.6
4	89.5	31	86.3	58	81.6
5	91.0	32	80.7	59	77.9
6	82.9	33	83.9	60	92.3
7	81.5	34	77.6	61	93.0
8	93.0	35	84.8	62	83.6
9	92.2	36	78.8	63	89.7
10	79.5	37	90.7	64	77.8
11	79.8	38	92.1	65	88.4
12	81.0	39	78.4	66	81.7
13	89.5	40	92.8	67	93.7
14	89.4	41	77.3	68	76.3
15	91.9	42	82.3	69	79.6
16	79.5	43	79.1	70	79.2
17	77.8	44	91.6	71	81.0
18	88.3	45	78.9	72	80.3
19	80.5	46	88.2	73	91.6
20	93.3	47	78.2	74	91.8
21	90.7	48	83.6	75	91.6
22	93.3	49	91.9	76	93.1
23	86.9	50	92.2	77	89.0
24	82.0	51	80.2	78	80.0
25	80.1	52	84.2	79	81.0
26	92.2	53	79.2	80	78.3
27	90.3	54	91.2		

These threshold parameters affect the size of the final decision tree branch. The greater the α value, the smaller the branch depth of the tree, while the greater the β value, the greater the branch depth of the tree. The setting of α and β has a great influence on the classification performance of the decision tree. Therefore, in order to achieve the best classification performance, the most appropriate threshold must be selected. After many repeated steps, the appropriate threshold is finally obtained, and the decision tree is determined by trying different values of α and β.

4. English Speech Feature Recognition System Based on Digital Means

From the perspective of service design, English speech feature recognition system needs to consider the service value of the system in addition to users, products, interactive behaviors, and scenarios. Therefore, English speech feature recognition is a system composed of five basic elements: person, process, object, product use scene, and service value. The English speech feature recognition system is designed around the above five basic elements. Its purpose is to coordinate the relationship between them and analyze their functions and properties to optimize the user experience. The system components are shown in Figure 1.

The process primarily relates to the actions initiated by the user when using the product in the service system, as well as the product's feedback behavior to the user. It also encompasses the interactions between users and other stakeholders. Good behavior may pique users' attention, make it more easy, quick, and pleasant to use, decrease errors, and elicit a more positive emotional response. The interactive behavior route of the English voice feature recognition system based on digital means is shown in Figure 2.

Internet of Things technology is applied to the English speech feature recognition system based on digital means. The three main characteristics of the Internet of Things are comprehensive perception, reliable transmission, and

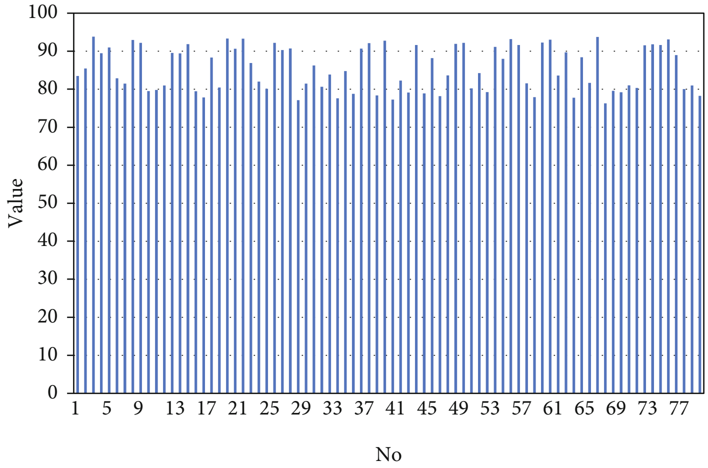


Figure 9: Statistical table of the accuracy of English feature recognition.

intelligent processing. From the perspective of the entire architecture of the Internet of Things, the Internet of Things consists of three major parts, from the low-end to the top-level are the perception layer, the network layer, and the application layer. The functional attributes carried by each layer are different. The perception layer uses various sensors to collect environmental data. The network layer is composed of mobile communication and the Internet, forming the bottom-end data transmission channel, and is responsible for transmitting the information collected by the sensors to the application layer, as shown in Figure 3.

Artificial intelligence-related technologies have entered an era of exponential expansion, thanks to the development and maturity of technologies such as cloud computing and big data. This affects not only just established industries' innovation and transformation but also people's everyday lives, bringing new behavioral experiences to human-computer contact. Intelligent speech technology requires expertise in a variety of fields, including acoustics, cognition, pattern recognition, and artificial intelligence. The intelligent voice system framework is made up of five parts. Accepting the user's voice input and turning it to text for the natural language understanding module is the responsibility of the voice recognition module. The natural language comprehension module inserts specified phrases into the conversation management module after comprehending the semantics of user input. The dialogue management module is in charge of coordinating calls from other modules, keeping track of the current conversation state, and passing the particular reply method to the natural language generation module for processing. The voice synthesis module receives a specified reply text from the natural language generation module. The voice synthesis module is in charge of converting the text into speech for the user. Figure 4 depicts the intelligent voice system framework.

Intelligent voice brings users a new way of interaction, which is more flexible, natural, and efficient than graphical interaction. The difference in the way of information transmission makes the elements of graphic interaction design and voice interaction extremely different. Graphic interaction design mainly takes six elements of color, font, motion, material, layout, and shape as the main design objects, while the voice interaction design needs to use the timbre of intelligent voice as the object, as shown in Figure 5.

Figure 6 shows a key figure map, which is a chart or map that shows the systematic portrayal of each key figure and the link between them. It may assist designers in swiftly determining the link between each position in the project in order to identify and solve difficulties in their interactive scenarios and assist in the subsequent optimization of the experience. Producers and consumers no longer have an interest connection once users buy things in the classic sense. Producers have transformed into service providers as product innovation methods have changed, and intangible services and physical goods are now linked and interconnected.

In the process of service design, the systematic innovation method, from the overall perspective, comprehensively considers the elements of human, service, and environment as well as the relationship between the elements and then reasonably plans the combination order and cooperation degree of the elements in the system, so as to maximize the performance of the whole service system. Based on the system innovation, it analyzed the scattered problem points in the intelligent voice service, as well as many stakeholders that may be involved in the service system, and reestablishes the task relationship model. On the one hand, it is needed to pay attention to the pain points of users' experience in the whole process of using intelligent voice service, which is the key element of service optimization. On the other hand, people, products, and environment elements in the intelligent voice service system need to be regarded as an interactive, interdependent, and comprehensive system with specific goals, so as to maximize the service of home English speech recognition system.

The whole system consists of five parts: PC, PCI adapter card, main switch, branch switch, and student terminal. The block diagram of the whole system is shown in Figure 7.

5. Performance Test of English Speech Feature Recognition System Based on Digital Means

Experimental study is used to evaluate the performance of the English voice feature recognition system based on digital techniques developed in this work. This paper's method mostly turns English speech into an electronic form before recognizing it. As a result, this work focuses on the impact of digital conversion of English speech and the recognition effect of English speech characteristics in the experimental investigation. First and foremost, the statistics of the digital conversion impact of English speech of the system created in this research are presented in this paper. We utilize the algorithm to identify 80 groups of student talks and transfer them to a digital form. Table 1 and Figure 8 illustrate the outcomes.

Through the above analysis, it can be seen that the English speech feature recognition system based on digital means constructed in this paper has a certain effect on the digital conversion of English speech. After that, it analyzed the accuracy of the English feature recognition of the system in this paper. The results obtained are shown in Table 2 and Figure 9.

From the above research, it can be seen that the English speech feature recognition system based on digitized segments constructed in this paper has certain effects and can play a certain role in intelligent English teaching.

6. Conclusion

English speech feature recognition plays an important role in supporting English learning. From a practical point of view, the improvement of English speech feature recognition through digital means can effectively improve the effect of English learning. It combined the actual needs of English speech feature recognition to apply digital means to English speech recognition. The reliability brought by digitization can effectively improve the ability of the chaotic system to resist channel interference and channel distortion, especially in its antiattack and interception capabilities. In this paper, time frequency analysis of chaotic signals and speech signals is carried out to eliminate noise in English speech features and improve the recognition effect of English speech features. Meanwhile, this paper constructs an English speech feature recognition system based on digital means to test the system's performance. The research results show that the method proposed in this paper has a certain effect.

Data Availability

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

Disclosure

A preprint has previously been published [23].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

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

[2] N. Q. K. Duong and H. T. Duong, "A review of audio features and statistical models exploited for voice pattern design," *Computer Science*, vol. 3, no. 2, pp. 36–39, 2015.

[3] M. Sarria-Paja, M. Senoussaoui, and T. H. Falk, "The effects of whispered speech on state-of-the-art voice based biometrics systems," *Canadian Conference on Electrical and Computer Engineering*, vol. 2015, no. 1, pp. 1254–1259, 2015.

[4] 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.

[5] T. Haderlein, M. Döllinger, V. Matoušek, and E. Nöth, "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, 2016.

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

[7] 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.

[8] 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.

[9] M. Woźniak and D. Połap, "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.

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

[11] 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, pp. 2302–2302, 2015.

[12] T. G. Kang and N. S. Kim, "DNN-based voice activity detection with multi-task learning," *IEICE Transactions on Information & Systems*, vol. E99.D, no. 2, pp. 550–553, 2016.

[13] H. N. Choi, S. W. Byun, and S. P. Lee, "Discriminative feature vector selection for emotion classification based on speech," *Transactions of the Korean Institute of Electrical Engineers*, vol. 64, no. 9, pp. 1363–1368, 2015.

[14] C. T. Herbst, S. Hertegard, D. Zangger-Borch, and P. Å. Lindestad, "Freddie mercury—acoustic analysis of speaking fundamental frequency, vibrato, and subharmonics," *Logopedics Phoniatrics Vocology*, vol. 42, no. 1, pp. 29–38, 2017.

[15] J. Al-Tamimi, "Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: implications for formal representations," *Laboratory Phonology*, vol. 8, no. 1, pp. 28–40, 2017.

[16] 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.

[17] 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.

[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] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[23] Y. Miao, Y. Huang, and Z. Da, *English speech feature recognition based on digital means*, Research Square, 2021.