A Self-Adaption Recognition Method for English Speech on Cloud Platform Based on Wavelet Neural Network

Xiaoqin Shi[a], Jinrong Wang[b]

College of Languages & Culture, Northwest A&F University, Yangling, Shaanxi, 712100, China

[a]1135014997@qq.com; [b]358320903@qq.com

Abstract. Although there are various recognition methods for digital English speech (DES) on cloud platform network, the accuracy of them is low and the average time consumption is too long. Due to above shortages, this study put forward a self-adaption recognition method based on wavelet neural network (WNN). At first, the study used zero-crossing rate as feature of voiceless and voiced sound of the DES and used WNN to carry out conversion for the English speech information. Then, the study obtained wavelet-scaling function and parameterized wavelet function via zero-pole model. The parameterized wavelet function was used as the feature vector of the recognition. Wavelet function base of DES feature was generated via stretching and translating transformation. The transformed wavelet function was used to build WNN model. Moreover, error entropy function of the recognition was calculated by introducing momentum factor and partial derivative of the error entropy function to adjust parameter of the built WNN model. Thus, the study achieved the self-adaption recognition of DES. Simulation results show that the proposed method has high recognition accuracy and short recognition time.

Keywords: Cloud platform, WNN, DES, Self-adaption recognition.

1. Introduction

The research on DES recognition on cloud platform network began in the early 1950s. With the continuous expansion and deepening of the research work, people have put forward higher and higher requirements for DES recognition. The ultimate goal of the DES recognition on cloud platform is to realize the free human-machine interaction as smooth as face-to-face communication between people. To some extent, it is to endow computers and other devices with hearing, so that machines can recognize human language, distinguish conversation content or speakers, and correctly convert human speech into written words or symbols with special significance that can be recognized by both machines and human[1-2], Furthermore, it is to control the machine through human will, let the machine assist English teaching, and liberate teachers and other staff from heavy work. According to relevant research data, the DES recognition on cloud platform will become another step in the transformation of human-computer interaction interface after the keyboard and mouse [3].

In recent years, there have been many methods for the studies of DES recognition on cloud platform. The more representative is the HMM (Hidden Markov Model) method which uses the logarithmic minimum mean square error to improve the signal-to-noise ratio of the collected noisy DES in the cloud platform network, like Chinese scholar Wang used the HMM in his study [4]; The intelligibility of DES is improved by using Wiener filter to remove the noise residue of DES; HMM is constructed to extract and recognize the features of the denoised DES. Chen [5] proposed a cloud platform network DES recognition method based on domain correlation feature transformation and fusion. This method uses domain correlation feature transformation to transform and generalize the extracted short-term spectral structure features and envelope features of DES; Different time granularities are used to divide DES into multiple levels, and SVM (Support Vector Machine) classifier is used to extract and recognize the features of DES after multi-level division. For example, Jiao [6] used SVM model in his study to classify and detect the errors in English pronunciation. Deng [7] proposed a method for DES recognition on cloud platform network based on deep neural network. This method estimates the parameters of the speaker and the speaking environment by constructing a Gaussian mixture model, which takes the parameter estimation results as the long-term characteristics

of DES, and inputs the estimated long-term characteristics of DES into the deep neural network for training. It can realize the cloud platform network DES recognition.

In view of the shortcomings that the above methods cannot achieve high accuracy and low time consumption for DES recognition on cloud platform, the study proposes an adaptive recognition method of DES on cloud platform based on WNN.

2. DES Recognition Principle

This principle obtains the linear spectrum of DES by performing discrete Fourier transform on the preprocessed cloud platform network DES sequence. The logarithmic spectrum of DES is obtained by using the logarithmic energy processing method, which was used in Wang's study [8]. The Mel-Frequency Cepstral Coefficients (MFCC) of DES can be obtained by discrete pre-transformation of the calculated logarithmic spectrum of DES. The MFCC of the first 12 dimensions is selected as the feature of DES. DES on cloud platform network is recognized via using linear prediction system, like the method which Liang used in his study [9].

Supposed $x(n)$ represents time domain signal of each English speech frame that the original DES, $s(n)$ on cloud platform network was preprocessed on emphasis, framing, windowing; and then, complement several zeros following $x(n)$ to form a speech sequence with length N, and the linear spectrum $X(k)$ of DES can be obtained by line discrete Fourier transform. The calculation formula is as follows:

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi nk/N},0\le n,k\le N-1 \quad (1)$$

The study input the Mel frequency filter bank (composed of several band-pass filters set within the English speech frequency range) to the above calculation results, $X(k)$, to obtain the Mel spectrum of the DES on the cloud platform network. The logarithmic energy processing method can be used to obtain logarithmic spectrum $S(m)$ of DES, and the calculation formula is:

$$S(m)=\ln\left(\sum_{n=0}^{N-1}\left|X(k)\right|^2 H_m(k)\right),0\le m<M \quad (2)$$

Where $H_m(k)$ represents the transfer function of band-pass filter set within the English speech frequency range; m is a constant which is bigger than or equals to 0; M means audio in English Number of band-pass filters set within the rate range.

The MFCC, $c(n)$, can be obtained by discrete pre transformation of the logarithmic spectrum, $S(m)$, of the DES calculated by the above formula (2), and $c(n)$'s calculation formula is as follows:

$$c(n)=\sum_{m=1}^{M-1}S(m)\cos\left(\frac{\pi n(m+1/2)}{M}\right),0\le m<M \quad (3)$$

According to the above calculation, the first 12 dimensions of $c(n)$ are selected as DES features which are used for subsequent identification, like what Yin used in his study [10].

Supposing the extracted DES feature is a noisy sequence, represented by $X(i)$, and its calculation formula is as follows:

$$X(i)=x(n)+N(i) \quad (4)$$

where $N(i)$ indicates that the DES feature sequence, which contains stationary random noise on cloud platform network.

Supposing $B_{j'}, j'=1,2,\cdots,p_1$ indicates that DES is linear prediction coefficient where voice is silent. The study used $B_{j'}$ to analyze the noisy sequence of DES features. After the analysis of features, DES can be recognized on cloud platform network. The calculation formula is as follows,

$$X(i)=\sum_{j'=1}^{p_1}B_{j'}X(i-j')=\sum_{j'=1}^{p_1}B_{j'}x(n-j')+\sum_{j'=1}^{p_1}B_{j'}N(i-j') \quad (5)$$

3. Method

3.1 Feature extraction of DES on cloud platform network

The zero-crossing rate is used as the feature of unvoiced or voiced DES. The wavelet function is used to transform the cloud platform network DES, and the wavelet scaling function was gotten, like what Qin did in his study [11], and the zero-pole model is used to parameterize the wavelet function of the digitized DES on the cloud platform network, and the parameterized wavelet parameters are used as the feature vector of DES recognition, which was used by many scholars, like Song [12].

Assumed $s_\omega(n)$ indicates windowed-digitalized DES in cloud platform network, where, ω Represents windowing coefficient τ Represents the autocorrelation coefficient; $R_\omega(\tau)$ Indicates the autocorrelation function of $s_\omega(n)$ which is calculated as follows,

$$R_\omega(\tau)=\sum_{n=-\infty}^{\infty} s_\omega(n)s_\omega(n-\tau)=\sum_{n=-\infty}^{N-\tau-1} s_\omega(n)s_\omega(n+\tau) \quad (6)$$

In general, during human-computer interaction, if the voice of a person is voiceless, its autocorrelation function of sound is similar to that of noise, while the spikes of the spectrum of DES are obvious in the voiced autocorrelation function. These spikes have periodic characteristics, and the periodicity of spikes is consistent with the periodicity of the original DES. It is assumed that Z_0 represents the short-time zero crossing rate of the original DES $s(n)$ on network platform, and it is the number of times that the voice signal waveform intersects the abscissa when a segment of English voice is gotten in network cloud platform , and the calculation formula is as follows

$$Z_0=\frac{1}{2}\left\{\sum_{n=0}^{N-1}\left|\mathrm{sgn}\left[s_\omega(n)\right]-\mathrm{sgn}\left[s_\omega(n-1)\right]\right|\right\} (7)$$

According to the above calculation, the unvoiced segment signal of DES has random characteristics and it will frequently cross the zero point. The short-time zero crossing rate of the unvoiced segment signal is usually high, while the short-time zero crossing rate of the voiced end signal is relatively low. Because the English pronunciation has an obvious difference between the short-time zero crossing rate of the voiced segment and the unvoiced segment of the DES, the zero-crossing rate can be used as a feature of voiced and voiced DES. In addition, the original feature extraction method is only applicable to the stationary English speech signals, but the network digitized DES of the cloud platform is a non-stationary signal, while the study used the wavelet function to transform the voice information and get the wavelet scale function as follows,

$$\varphi(\omega)=e^{-j'\frac{\omega}{2}}\left(\frac{\sin\frac{\omega}{2}Z_0}{R_\omega(\tau)\frac{\omega}{2}}\right)^{-3} \quad \ldots (8)$$

According to the above calculation results, the zero-pole model is used to parameterize the wavelet function of DES in cloud platform, and the parameterized wavelet function is used as the feature vector of DES recognition, whose calculation formula is as follows,

$$L(\omega)=e^{-j'\frac{\omega}{2}}\cos^3\varphi(\omega) \quad (9)$$

3.2 Information adaptive recognition based on WNN

The wavelet function base is gotten by telescopic translation transformation of cloud platform network digitized DES and the cloud platform network l wavelet neural network model has been built by using the transformed wavelet function base, and then, the error entropy function of English speech recognition is computed. The parameters of the model are adjusted by using the partial derivative pair of momentum factor and error entropy function [13]. So, the adaptive recognition of DES on cloud platform network can be realized by adjusted parameters.

It is an assumptions that $\psi(y)$ represents a mother wavelet function of information features of the above extracted English voice on cloud platform, which generates a group of small wave function basis, and the calculation formula is as follows, Where, a and b represent the telescopic scale factor and the translational scale factor respectively.

$$\psi_{a,b}(y)=\frac{1}{\sqrt{a}}\psi(\frac{y-b}{a})\cdot L(\omega) \quad (10)$$

y_κ Represents the input sample of the κ th voice feature in the input layer of wavelet neural network; $\gamma_{i'}$ indicates the i' - th output voice in the output layer of the network ; $\omega_{i''j''}$ refers to the weight of connecting the output layer node and the hidden layer node of wavelet neural network; $V_{j''\kappa}$ Represents the connection between hidden layer node and input layer node of wavelet neural network; $\omega_{i''0}$ and $\omega_{j''0}$ indicate the thresholds of the i'' -th output layer node and the j'' -th node of the network hidden layer; P refers to the number of patterns of input DES feature sample Items; σ Represents the sigmoid function; m' 、 n' 、 g denote the number of the nodes of input layer, hidden layer, and output layer of wavelet neural network output, respectively; t is the sampling time; η represents network learning rate, and the l wavelet neural network model of cloud platform network can be built according to the above parameters. The calculation formula is as follows,

$$y_\kappa(t)=\sigma\left|\sum_{j''=0}^{n'}\omega_{i''j''}\psi_{a,b}(y)\left|\sum_{\kappa=0}^{m'}V_{j''\kappa}\omega_{i''0}\omega_{j''0}\eta\right|\right|, i''=1,2,\mathrm{L}\ ,g\ \ldots \quad (11)$$

4. Calculating error entropy function of DES recognition

y_P^κ represents the P -th of the Input mode of cloud platform network DES; $y_{i''}^P$ and $d_{i''}^P$ respectively represent the actual output and expected output of the i'' -th wavelet neural network of the P -th input mode of voice information, and then the error entropy function of information recognition can be obtained, and its formula is

$$E(t)=-\sum_{P=1}^{P}\sum_{i''=1}^{g}\left|d_{i''}^P\ln y_{i''}^P+(1-d_{i''}^P)\ln(y_\kappa(t))\right| \quad (12)$$

3.3 Parameter adjustment of WNN model

The parameters of the above wavelet neural network model can be adjusted by introducing the momentum factor μ, and the digitized DES in cloud platform network can be recognized by using adjusted parameters of wavelet neural network model. The formulas of adjusted parameters are as follows

$$V_{j''\kappa}(t+1)=V_{j''\kappa}(t)-\eta\frac{\partial E(t)}{\partial V_{j''\kappa}}+\mu V_{j''\kappa} \quad (13)$$

$$\omega_{i''j''}(t+1)=\omega_{i''j''}(t)-\eta\frac{\partial E(t)}{\partial \omega_{i''j''}}+\mu\omega_{i''j''} \quad (14)$$

$$a(t+1)=a(t)-\eta\frac{\partial E(t)}{\partial a}+\mu a \quad (15)$$

$$b(t+1)=b(t)-\eta\frac{\partial E(t)}{\partial b}+\mu b \quad (16)$$

$\frac{\partial E(t)}{\partial V_{j''\kappa}}$ 、 $\frac{\partial E(t)}{\partial \omega_{i''j''}}$ 、 $\frac{\partial E(t)}{\partial a}$ and $\frac{\partial E(t)}{\partial b}$ respectively represent the error partial derivatives of the parameters of the wavelet neural network model.

3.4 Simulation test and result analysis

Two types of English speech databases in the cloud platform network are selected as samples. The test data are the native English database and the Chinglish database. Among them, the native English database contains the continuous English pronunciation of 200 people for a total of 60 hours. The

experimental test set contains 1500 English short sentences (each English short sentence contains 1 ~ 2 words). Three 100 English short sentences are randomly selected from the test machine as the experimental test set; Chinglish data includes 30 hours of continuous English pronunciation of 50 men and 50 women (150 English pronunciation short sentences per person, including 140 English short sentences as the training data set and the remaining 10 as the recognition test set.)

As shown in Figure 1 and Figure 2, the pronunciation waveforms of English native speakers and Chinese pronunciation waveforms are given respectively when the same English short sentence are spoken without interference.

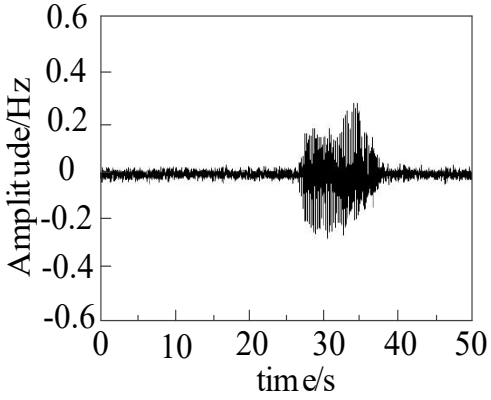

Fig. 1. Native speakers' pronunciation waveform

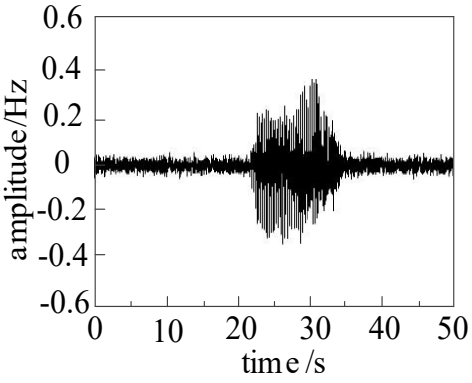

Fig. 2. Chinese speakers' pronunciation waveform

The study injected 10%, 20%, 30% and 50% random noise interference into Fig. 1 and Fig. 2 respectively and, then, the study used the methods Wang, Chen, Deng, as well as the adaptive recognition method which the study proposed based on wavelet neural network to recognize Fig. 1 and Fig. 2 after injecting random noise interference. The recognition results are shown in Table 1. ①, ②, ③ and ④ in Table 1 respectively represent the method of this study, the method of Chen [5], the method of Deng [7], and the method of Wang [4]; Noise refers to the proportion of random interference noise added; AC stands for recognition accuracy.

Table 1. Results of DES recognition under different methods

method	noise%	①	②	③	④
		AC%			
figure 1	10%	99.2	87.3	90.1	77.6
	20%	98.7	82.6	88.5	71.9
	30%	97.6	77.5	84.7	66.7
	50%	96.5	73.2	79.2	63.2

figure 2	10%	98.9	81.4	88.4	72.8
	20%	97.3	76.3	84.3	64.7
	30%	96.1	72.9	78.2	61.5
	50%	95.8	68.4	75.1	49.1

From the experimental results in Table 1, it can be seen that the recognition accuracy of the method in Wang's study is the lowest, especially when information recognition is carried out for the pronunciation of native English speakers with random noise, with the increasing proportion of random noise, the recognition accuracy becomes lower and lower. When the proportion of random noise injected reaches 50%, the recognition accuracy of the method in Wang's study is lesson than 50% (49.1%) for Chinglish pronunciation; The method of Deng and the method of Chen have relatively high recognition accuracy when recognizing the pronunciation information of native English speaker, but the recognition accuracy also drops sharply when recognizing the pronunciation information of Chinglish short sentences with random noise, which shows that the three comparison methods cannot effectively improve the recognition accuracy of Chinglish phrase information, but the recognition accuracy of the proposed method is high. Especially under the condition of low signal-to-noise ratio, the effect is more remarkable. At the same time, compared with other three methods, the proposed method effectively improves the accuracy of Chinglish phrase information recognition. This is because the WNN model constructed by the proposed method can adapt to different speech environments by adaptively adjusting the model parameters.

According to the above experimental data, 100, 300, 500, 800 and 1000 native English sentences and Chinese English sentences are recognized by using the methods of this study, Chen, Deng and Wang. The average recognition time is counted. The results are shown in Table 2. I and II in Table 2 represent native English pronunciation and Chinglish pronunciation respectively.

Table 2. Comparison of average recognition time under different methods

Method	noise%	①	②	③	④
		Time/s			
I	100	0.06	0.51	0.48	0.24
	300	0.08	0.62	0.56	0.36
	500	0.11	0.78	0.79	0.49
	800	0.15	1.26	1.23	0.71
	1000	0.18	1.69	1.58	0.98
II	100	0.07	0.73	0.69	0.49
	300	0.09	0.84	0.72	0.68
	500	0.13	1.96	1.85	0.95
	800	0.21	2.48	2.69	1.47
	1000	0.26	3.66	3.52	2.31

By observing the experimental results in Table 2, it can be found that the average recognition time of the proposed method is the shortest, and does not increase significantly with the increase of the number of English short sentences to be recognized. At the same time, there is no significant difference between the recording of English short sentences by native speakers or Chinese English speakers. The average recognition time of the method in Wang's study is also short, second only to the proposed method, followed by the method of Deng and the method of Chen. Based on the analysis

of the experimental results in Table 2 above, the average recognition time of the method of Wang is short, but the recognition accuracy is low, which time consumption is improved at the cost of sacrificing recognition accuracy. The other two comparison methods are far inferior to the proposed method in terms of recognition accuracy and average recognition time.

4. Conclusion

In order to better realize human-computer interaction, the current DES recognition methods are studied for computer-aided English education and teaching. Aiming at solving the problems of low accuracy and long time-consumption of pronunciation information recognition, especially for Chinglish pronunciation recognition, the study proposed a cloud platform network DES adaptive recognition method based on WNN, and its effectiveness is proved by simulation and comparison. The proposed method improves the accuracy and reduces time consumption of DES on cloud platform network, compared with other current DES recognition methods. The study will give some suggestion for further study of rating students' pronunciation by machines in spoken English test.

Acknowledgements

Firstly, the study has been approved by the project "The Study of the Construction of Oral CAF Evaluation System Based on English Corpus" (No. XGH21053) from Shaanxi Higher Education Society, by the teaching reform project, "The Study of the Optimization and Evaluation of Oral CAF Indexes Based on English Corpus" (No. JY2103201) from Northwest A&F University, and by the experimental technique project "Research on the Function Innovation of Oral Test Based on Digital Speech Technology" (No. SY20220217) from Northwest A&F University. The author thanks the departmental Office for the support. Secondly, the author also thanks co-workers for their helps in the study. At last but not least, the authors appreciate what assistants and other persons have done for the study.

References

[1] Qin, Ch., Zhang L. (2017) Feature Extraction for low-resource speech recognition based on DNN. Acta Automatica Sinica, 43(7), pp:1208-1219. http://doi.org/10.16383/j.aas.2017.c150654.

[2] Wang, Y., & Zhao, P. (2020). A Probe into Spoken English Recognition in English Education Based on Computer-Aided Comprehensive Analysis. International Journal of Emerging Technologies in Learning (iJET), 15(03), pp. 223–233. https://doi.org/10.3991/ijet.v15i03.12937.

[3] Zhou, F., Yu, Y.. (2017). Application of Group Delay Spectrum Parameters in Chinese. Speech Recognition. Signal Processing, 33(9):1215-1220. http://doi.org/10.16798/j.issn.1003-0530.2017.09.008.

[4] Wang, Q., Zeng, Q. N., Xie, X. M.. (2017). Research on speech recognition in low SNR environment. Acoustic Technology, 36(1):50-56. http://dio.org/10.16300/j.cnki.1000-3630.2017.01.010.

[5] Chen, B., Hu, P.G., Qu, D.. (2016). Speech Recognition Based on Correlation Feature Transformation and Fusion in Subspace Domain. Journal of Xi 'an Jiaotong University, 50(4), pp. 60-67.

[6] Jiao, F., Song, J., Zhao, X., Zhao, P., & Wang, R. (2021). A Spoken English Teaching System Based on Speech Recognition and Machine Learning. International Journal of Emerging Technologies in Learning (iJET), 16(14), pp. 68–82. https://doi.org/10.3991/ijet.v16i14.24049

[7] Deng, K., Ou, Zh.. (2016). Speech recognition method based on subspace domain correlation feature transformation and fusion. Application Research of Computers, 33(7), pp. 1966-1970.

[8] Wang, H., Wu D., Liu, J.. (2017). Automatic Speech Recognition Based on Time-domain Modeling. Computer Engineering and Applications, 53(20), pp. 243-248.

[9] Liang, Y., Yang, P., Sun, H.. (2017). The optimization and simulation of Robot's real-time recognition for special people speech. Computer Simulation, 34(10), pp. 286-290.

[10] Yin, Z. (2018). Training & Evaluation System of Intelligent Oral Phonics Based on Speech Recognition Technology. International Journal of Emerging Technologies in Learning (iJET), 13(04), pp. 45–57. https://doi.org/10.3991/ijet.v13i04.8469.

[11] Qin, Ch., Zhang, L.. (2016). Acoustic Modeling based on Convolutional Neural Network with Multi-stream Features for Low-resource Speech Recognition. Journal of Computer Applications, 36(9):2609-2615.

[12] Song, Q. S., Tian, Zh. X., Sun, W., et. al.. (2016) A Combined Dimension Reduction Method for Isolated l Speech Recognition [J]. Journal of Xi 'an Jiaotong University, 50(6), pp. 42-46.

[13] Xiang, B., Jing, X., Yang, H.. (2017). Vehicle-mounted Speech Recognition Based on Noise Classification and Compensation. Computer Engineering, 43(3), pp. 220-224.