Received: November 14, 2017
Revision received: April 2, 2018
Accepted: April 12, 2018

www.estp.com.tr
DOI 10.12738/estp.2018.5.033 ♦ October 2018 ♦ 18(5) ♦ 1341–1350

Research Article

Using Computer Speech Recognition Technology to Evaluate Spoken English

Yanping Zhang[1]
Yulin University

Limin Liu[2]
Huzhou University

Abstract

Spoken English learning has always been the largest obstacle to the English communicative competence of Chinese students. This paper adopts the in-depth learning method based on computer-aided speech recognition technology to achieve the automatic speech recognition of spoken English and to overcome the defects of spoken speech evaluation in the traditional mode such as fairly high subjectivity and low evaluation efficiency. Meanwhile, the evaluation method of the traditional computer-aided spoken speech quality has been improved, and multi-parameter indicators such as accuracy, speed, rhythm and intonation have been taken into account. Through the establishment of the objective and efficient spoken English speech recognition and evaluation model, we are able to provide true and credible evaluations and timely feedback guidance for learners. The research results effectively guide learners to study spoken English, which uplifts the existing level of spoken English learning in China.

Keywords

Computer Speech Recognition • in-depth Learning Method • Multi-Parameter Index • Speech Recognition and Evaluation Model • Spoken English

[1]**Correspondence to:** Yanping Zhang (A.P.), School of Foreign Languages, Yulin University, Yulin 719000, China. Email: xyzruby@sina.com
[2] IT School, Huzhou University, Huzhou 313000, China. Email: llm@zjhu.edu.cn

Citation: Zhang, Y. P., Liu, L. M. (2018). Using Computer Speech Recognition Technology to Evaluate Spoken English. *Educational Sciences: Theory & Practice, 18*(5), 1341-1350. http://dx.doi.org/10.12738/estp.2018.5.033

With the process of globalization, English has been increasingly widely used as an international language. Chinese English learners have encountered great difficulties in learning spoken English, and their shortcomings become more and more prominent. Domestic educational institutions have introduced different approaches to improve Chinese students' spoken English capabilities. However, restricted by educational resources, educational environment and other factors, the teaching results of spoken English are still ineffective. A number of English learners have a "Chinglish" accent (Gonçalves *et al*., 2004), and they are rebuffed in the actual English communications. The rapid development of the learning technology assisted by computer language offers a new way to spoken English learning, whose core is speech recognition technology and speech evaluation technology, while the former is the key. Speech recognition technology, also known as automatic speech recognition (ASR), refers to the technology of converting speech signals into corresponding commands or texts by means of automatic identification and understanding of machines (generally refers to computers) (Gerard *et al*., 2016; Wang *et al*., 2012), thereby achieving intelligent speech interaction between humans and machines. Therefore, speech recognition has become a popular research topic in recent years. Due to the complex pronunciation variation, the substantial amount of speech signal data, the high dimensional speech feature parameters, the large quantity of calculations of speech recognition and evaluation, a substantial amount of speech signal processing requires hardware and software resources and algorithms to reach a higher level (Espy, 2005). With the development of in-depth learning, big data and cloud computing technology, speech recognition and evaluation techniques have also evolved rapidly (Cooke *et al*., 2006). The study on English speech recognition based on in-depth learning greatly improves the processing ability of speech information, enhances the users' efficiency of information acquisition, and provides better user experience.

This paper first briefly introduces the process of speech processing, and then elaborates the basic idea and the training content of in-depth learning method. Through the interpretation and the analysis on multi-parameter indicators, a speech recognition and evaluation model for English spoken language is established. In the end, experiments are conducted to verify that this model has high credibility, provides timely feedback on the pronunciation of spoken English learners, and locates the difference between their pronunciation and standard pronunciation. Besides, an objective and efficient evaluation model enhances the learning efficiency. Therefore, computer-aided speech recognition technology has a positive effect on the improvement of spoken English.

Introduction of speech signal preprocessing technology

Speech recognition process

Figure 1 illustrates the general process of speech recognition. This paper applies computer sound card to collect speech signal and transforms analogue quantity into digital quantity and the speech signal is extracted.

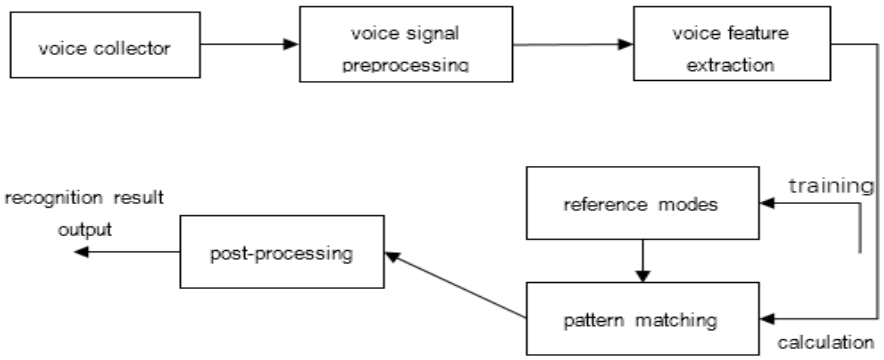

Figure 1. Language recognition process diagram.

According to the Nyquist sampling theorem (Song *et al*., 2012), when the sampling frequency is greater than twice the highest frequency of the signal, the sampling distortion is close to zero and the human speech frequency range is 40 to 4000Hz. Therefore, the sampling frequency in this paper is set as 8KHz, and the speech signal is preprocessed. The process of speech recognition is completed through matching patterns or selecting mode training with the feature parameters.

Speech signal preprocessing

Preprocessing operations such as pre-emphasis, framing, windowing, and endpoint detection must be performed before the speech signal is analyzed and processed (Thomas and Ravindran, 1971). The purpose of these operations is to eliminate the effects made by high-order harmonic distortion, high-frequency, aliasing, and other factors on the quality of speech signals that are caused by human vocal organs and speech signal acquisition devices (Villchur, 1973). Speech preprocessing affects the result of speech feature extraction. Smooth and uniform speech signal provides parameters with better speech feature extraction quality, so as to improve the quality of speech processing.

Parameter extraction of speech feature

The extraction of speech feature parameter refers to the removal of irrelevant and redundant speech information. The original speech information has varying degrees of deviation because each person has a different accuracy, timbre, and range, etc. Therefore, each speech is required to perform feature parameter extraction. To a certain degree, the process of feature parameter extraction determines the performance of speech processing.

Meg frequency cepstrum coefficient (MFCC) is used to simulate the response of human ear to different frequency signals, as shown in Figure 2, which is the feature parameter extraction process of MFCC.

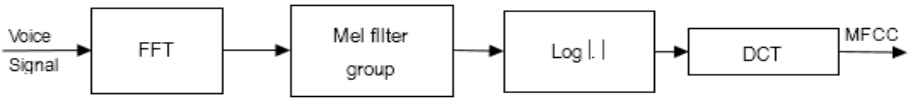

Figure 2. MFCC feature parameter extraction process diagram.

Through the Fourier transform of the speech signal, the filter is carried out by using the Mel filter. Furthermore, feature parameter extraction of the speech signal is completed by logarithmic operation and discrete cosine transform (DCT). The specific process is referred in the section of literature review (Furoh *et al*., 2013).

Multi-parameter speech quality evaluation based on in-depth learning method

Introduction to in-depth learning method

The concept of in-depth learning originates from the artificial neural network. As one kind of unsupervised learning, the in-depth learning structure contains a number of hidden layers (Xing *et al*., 2010). Hinton has proposed an in-depth learning algorithm for neutral network, an effective method of establishing a multi-layer neural network on unsupervised data. Consequently, it is possible to train at least seven layers of neural network, which is called in-depth neural network (Qi *et al*., 2012). In-depth learning combines the features of the low layers and forms more abstract high layers to represent its attribute categories or characteristics and to discover the distributed feature representation of the data.

At present, in-depth learning technology has been successfully applied in the issue of multi-pattern classification and verified in the field of speech recognition. Although this area is in its early stages of development, its development will undoubtedly exert a significant impact on artificial intelligence and machine learning (Christiansen, 2011). This paper applies in-depth learning method on English speech recognition and tries to improve its accuracy, speed and other indicators.

Advantages of computer-aided speech quality evaluation

In the traditional process of subjective English speech evaluation, there exists the shortcoming of the subjectivity. As the main body of the evaluation, experts or teachers are affected by emotions and other multiple factors, and their evaluations on the same speech in different periods may also differ.

In order to evaluate the speech quality more objectively, we apply computer technology on speech recognition and combine the in-depth learning method, which effectively overcomes the shortcoming caused by the subjectivity of artificial evaluation. Through the comparative analysis with the evaluation model, the input speech signal is analyzed objectively and then the scores of the evaluation indicators are given and evaluated by the computer.

Computer-aided evaluation indicators

Computer-based English pronunciation quality evaluation mainly carries out comprehensive analysis and evaluation on accuracy, rhythm, intonation and other aspects.

Accuracy evaluation. Accuracy evaluation mainly examines whether the content of the pronounced sentence is complete and accurate, whether the pronunciation is clear and fluent, and whether there is

mispronunciation (Hieronymus and Kadambe, 1997). This paper employs MFCC coefficient that is based on human ear auditory model as the evaluation parameter of the accuracy. A speech recognition model is constructed through the in-depth belief network to judge whether the content is complete and correct. Meanwhile, the correlation coefficient of MFCC characteristic between the standard sentence and the input sentence is calculated to judge whether the pronunciation is clear and fluent. These two factors are integrated to conduct accuracy evaluations and feedbacks on the quality of English speech.

Rhythm evaluation. English is a typical language that applies stress to express the meanings. The rhythm of each sentence and the structure of the tone determine the meanings of the sentence. Different rhythms of the sentence are reflected by the differences in accuracy, stress, length, degree of urgency and other factors of the language. By virtue of computer speech processing unit, we extract the stress part of the spoken English pronunciation, compare it with the extracted stress of the sample sentence and then give the scores (Zechner *et al*., 2009).

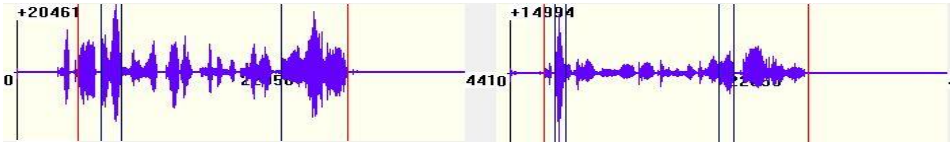

Figure 3. Accent extraction of sentence “It will be the place where we always put it”.

As shown in Figure 3, by means of computer-aided speech extraction software, the stress of the sentence “It will be the place where we always put it” pronounced by the subject is extracted to compare with that of the standard pronunciation.

In Figure 3, the right side shows the stress distribution of the standard pronunciation, and the left side illustrates the stress situation of the subject’s pronunciation. Through the further fitting evaluation by the computer, the evaluation scores of the stress rhythms of spoken English pronunciation are derived.

Intonation evaluation. Like the Chinese language, different English intonations represent different emotional needs of the speakers. Thus, intonation evaluation is also a key part of spoken English speech.

In this paper, the speech is segmented by computer speech recognition technology, and the accuracy in English sentences (Borde *et al*., 2015) is extracted by autocorrelation function (ACF). Similar to rhythm evaluation, through the comparison and the fitting degree with standard sentences, the accuracy of spoken English speech is evaluated and scored.

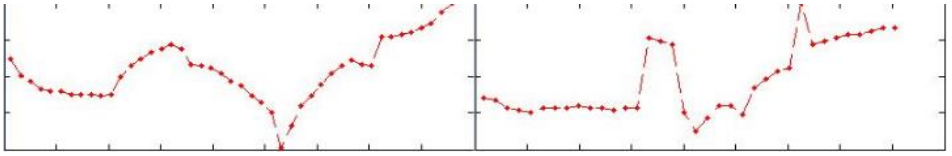

Figure 4. Intonation extraction of sentence “It will be the place where we always put it”.

In Figure 4, the right side demonstrates the accuracy distribution of the standard pronunciation, and the left side indicates the accuracy situation of the subjects' pronunciation. Through further feedback evaluation and analysis, the scores of the accuracy of the spoken English pronunciation are obtained.

Establish a multi-parameter speech quality evaluation model based on in-depth learning method

This paper takes the spoken English of college students as the object of study. On the basis of accuracy, speed, rhythm, intonation and other indicators of the speech quality evaluation, this paper conducts a comprehensive analysis on the relationship among the indicators, focuses on the weights of various indicators in the overall speech quality evaluation, establishes a multi-parameter English speech quality evaluation model and a method for college students, and comprehensively evaluates the speech quality in a reasonable and objective way (Kewleyport *et al*., 2002).

An empirical study on speech recognition in spoken English

Empirical research methods

This paper selects 24 college students, including 15 males and 9 females. Their speeches are recorded by the recording software to conduct language recognition processing. The sampling frequency is 16KHz. The sentences prepared for spoken English pronunciation are as follows:

1. It will be in the place where we always put it.
2. The clothes are lying on the fridge.
3. She will hand it in on Wednesday.
4. Tonight I could tell him.
5. The black sheet of paper is up there beside the piece of timber.
6. In seven hours the team leader will come.
7. What are the bags standing there under the table?
8. They have just carried it upstairs and now they are coming down again.
9. At the weekends I always go home and see Agnes.
10. I just want to take this away and then go for a drink with Karl.

A total of 10 different sentences and 240 different spoken English phonetic samples are collected. The speech data is preprocessed in the order described in the previous section.

Speech evaluation strategy of spoken English

The purpose of this experiment is to validate the performance of the multi-parameter evaluation model based on the in-depth learning method proposed in this paper. Besides, human evaluations are taken as a reference to evaluate its credibility. Three coefficients are introduced: agreement, adjacent agreement and Person interaction coefficient. These three coefficients are used to represent the consistent relationship between computer-based language recognition evaluation and human evaluation to indirectly evaluate the computer-based spoken English recognition model.

$A_{agreement}$ =(The number of samples of computer evaluation and human evaluation)/(The total number of samples) (1)

$A_{adjecent\ agreement}$ =(The number of agreeable samples of computer evaluation and human evaluation+the number of adjecent samples of computer evaluation and human evaluation)/(The total number of samples) (2)

Pearson correlation coefficient is a statistical measurement that reflects the linear correlation degree of the two variables, with a value range of [-1,1]. The greater the absolute value, the stronger the correlation is. Otherwise, the weaker it is.

$$r = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - (\sum x_i)^2}\sqrt{N\sum y_i^2 - (\sum y_i)^2}} \quad (3)$$

Analysis and discussion on experimental results

Based on the computer-based speech recognition technology, four parameters—accuracy, speed, rhythm and intonation are analyzed synthetically concerning 240 sentences of the students. The comparison results of the artificial recognition modes are illustrated in Table 1 and Table 2.

Table 1
Evaluation Results of the Experiment—the Number of Samples

Indicator	No diff	One level diff	Two level diff	Three level diff
Accuracy	207	32	1	0
Speed	197	43	0	0
Rhythm	204	33	3	0
Intonation	192	44	4	0

Table 2
Evaluation Results of the Experiment—Statistical Indicators

Indicator	Agreement	Adjacent Agreement	Person
Accuracy	86.25%	99.58%	0.8
Speed	82.08%	100%	0.493
Rhythm	85.00%	98.75%	0.543
Intonation	80.00%	98.33%	0.627

As indicated by Table 1, the difference between the majority of computer-aided speech recognition samples and human identification samples lies within Level 1 difference, and the number of consistent samples is large. In terms of statistical indicators, the agreement and the adjacent agreement between computer-aided speech recognition and human identification are higher than 80%, and Pearson correlation coefficient is also above 0.5. The four key elements of spoken English speech are accuracy, speed, rhythm and intonation. In light of this, this paper indicates that the computer-aided English spoken language recognition technology based on the proposed method has a high recognition rate and provides timely feedbacks to the students about the elements that need to be improved.

Next, in order to take into account the comprehensive impacts of the various evaluation indicators, SPSS software is applied to evaluate the values of the weights of the evaluation indicators, as shown in Figure 4:

$$\text{Score} = \text{Accuracy} \times 0.44 + \text{Speed} \times 0.106 + \text{Rhythm} \times 0.341 + \text{Intoation} \times 0.312 - 0.397 \quad (4)$$

The values in Table 1 and Table 2 are substituted into Formula 4 to obtain the variance analysis result: F=4.805 and P=0.003, which indicates that the scores have a significant relationship with accuracy, speed,

rhythm and intonation. In other words, the four indicators have a high accuracy for the overall evaluation on spoken English, thus ensuring the credibility of the experimental results.

Table 4 illustrates the overall evaluation on 240 sample sentences. According to Table 3, the overall agreement rate is 87.5%, the coincidence rate is 87.5% and Pearson correlation coefficient is 0.722.

Table 3
The Overall Evaluation Results

Indicator	Agreement	Adjacent agreement	Person
Accuracy	87.5%	100%	0.722

In the evaluation of English pronunciation quality, accuracy is the most essential indicator which requires accurate content, fluent pronunciation and no obvious pronunciation error. Rhythm and tone mainly express the speaker's emotions, enhance the speech tone, and make the speech closer to the real life. Speed varies from person to person. It is acceptable as long as it is not too fast or too slow to affect the listener's understanding, so the weight is relatively small.

The analysis on the empirical research process and the experimental results indicates that the computer-based speech recognition technology proposed in this paper holds favorable performance in the evaluation of spoken English speech, which can provide accurate and timely feedbacks for English learning, as well as theoretical significance for the development of spoken English education and learning.

Conclusion

The technology of speech recognition and speech evaluation is the key to computer-aided spoken English learning. In this paper, the in-depth learning method is applied to speech recognition, which improves the recognition rate and recognition efficiency. At the same time, the multi-parameter model is adopted to consider the evaluation indicators that affect spoken English pronunciation and to provide more convincing evaluation results. Besides, empirical analysis has been conducted to verify the practical applied effects of the evaluation model that builds on computing speech recognition technology on English spoken pronunciation. The following conclusions are drawn.

(1) Compared with the traditional evaluation on spoken English that emphasizes vocabulary and grammar, the multi-parameter evaluation model based on accuracy, speed, rhythm and intonation can reflect the level of spoken English more accurately.

(2) Through the comparative analysis on the manual evaluation results jointly conducted with several experts, the computer-aided speech recognition technology evaluation model adopted in this paper has high recognition accuracy.

(3) Computer-aided speech recognition technology exerts a positive impact on the evaluation and the improvement of spoken English. In China, spoken English education should actively adopt the computer-aided speech recognition technology to guide the learning and the teaching of spoken English.

References

Borde, P., Varpe, A., Manza, R., & Yannawar, P. (2015). Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition. *International Journal of Speech Technology*, *18*(2), 167-175. https://link.springer.com/article/10.1007/s10772-014-9257-1

Gerard, O., Carthel, C., & Coraluppi, S. (2016). Automatic detection and classification of beaked whale acoustic signals, *Traitement Du Signal*, *33*(1), 73-94. https://doi.org/10.3166/TS.33.73-94

Christiansen, T. (2011). Fluency and pronunciation in the assessment of grammatical accuracy in spoken production: An empirical study an empirical study. *Journal of Biological Chemistry*, *286*(39), 33795-33803.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, *120*(5-1), 2421-2424.

Espy, W. C. (2005). Automatic speech recognition. *Journal of the Acoustical Society of America*, *117*(4), 2403-2403.

Furoh, T., Fukumori, T., Nakayama, M., & Nishiura, T. (2013). Detection for Lombard speech with second-order Mel-frequency cepstral coefficient and spectral envelope in beginning of talking-speech. *Journal of the Acoustical Society of America*, *133*(1), 3246.

Gonçalves, J. P., Aluisio, S. M., Oliveira, L. H. M. D., & Jr, O. N. O. (2004). A learning environment for English for academic purposes based on adaptive tests and task-based systems. *Lecture Notes in Computer Science*, *3220*, 1-11.

Hieronymus, J., & Kadambe, S. (1997). Robust spoken language identification using large vocabulary speech recognition. *Journal of Membrane Science*, *2(1)*, 1111-1114. https://doi.org/10.1109/MASS.1995.528223

Kewleyport, D., Dalby, J., & Burleson, D. (2002). Speech intelligibility training using automatic speech recognition technology. *Journal of the Acoustical Society of America*, *112*(5), 2303. http://dx.doi.org/10.1121/1.4779274

Qi, Y., Dong, B., Ge, F., & Yan, Y. (2012). Text-independent pronunciation quality automatic assessment system for English retelling test. *Journal of the Acoustical Society of America*, *131*(4), 3234. http://dx.doi.org/10.1121/1.4708063

Song, Z., Liu, B., Pang, Y., Hou, C., & Li, X. (2012). An improved Nyquist–Shannon irregular sampling theorem from local averages. *IEEE Transactions on Information Theory*, *58*(9), 6093-6100. https://dx.doi.org/10.1109/TIT.2012.2199959

Thomas, I. B., & Ravindran, A. (1971). Preprocessing of an already noisy speech signal for intelligibility enhancement. *Journal of the Acoustical Society of America*, *49*(1), 133.

Villchur, E. (1973). Signal processing to improve speech intelligibility in perceptive deafness. *Journal of the Acoustical Society of America*, *53*(6), 1646-1657. https://dx.doi.org/10.1121/1.1913514

Wang, L., Qian, Y., Scott, M. R., & Chen, G. (2012). Computer-assisted audiovisual language learning. *Computer*, *45(6)*, 38-47. https://dx.doi.org/10.1109/MC.2012.152

Xing, Z., Yong, Q., Pang, X., Jia, L., & Yuan, Z. (2010). Modelling of the automatic depth control electrohydraulic system using RBF neural network and genetic algorithm. *Mathematical Problems in*

Engineering, 4, 242-256.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, *51*(10), 883-895. https://dx.doi.org/10.1016/j.specom.2009.04.009