Journal of Combinatorial Mathematics and Combinatorial Computing, 120: 275–283
DOI:10.61091/jcmcc120-024
http://www.combinatorialpress.com/jcmcc
Received 6 February 2024, Accepted 30 May 2024, Published 30 June 2024

Combinatorial Press

Article

Enhancing English Pronunciation Assessment in Computer-Assisted Language Learning for College Students

Guochang Zhang[1,*]

[1] School of Foreign Languages, Zhengzhou University of Industrial Technology, Zhengzhou 451150, Henan, China

* **Correspondence:** zhangguochang@zzuit.edu.cn

Abstract: The evolution of computer science and the innovations in language teaching methodologies have paved the way for computer-assisted language learning (CALL) technology to tackle pertinent challenges. While existing CALL systems primarily emphasize vocabulary and grammar acquisition, their evaluation mechanisms often rely on a limited set of criteria, resulting in a simplistic assessment of learners' pronunciation skills. This oversight underscores the need for a more comprehensive approach. In response, this study targets Chinese college students' English oral proficiency and aims to enhance the conventional computerized evaluation method. Our approach involves integrating multiple assessment parameters, including pitch, speed, rhythm, and intonation. For instance, pitch assessment is grounded on frequency central feature parameters, while speech speed evaluation considers speech duration, thus enriching the evaluation framework. Through experimental validation, the efficacy of our method in evaluating pitch, speed, rhythm, and intonation has been substantiated, reaffirming its reliability.

Keywords: Computer-assisted language learning (CALL), Language teaching methodologies, Oral proficiency, Evaluation method

1. Introduction

With global integration and the increasing internationalization of China, the demand for English learning in China is growing rapidly [1–3]. And the lack of good English teachers, and the traditional classroom teaching cannot meet the needs of English learning due to the constraints of time and location. The combination of all these reasons has made English teaching and learning a major problem for the people of China. English learning has also become one of the hot spots in the field of education. With the development of computer science and technology and advances in language teaching and learning methods, CALL technology has made it possible to solve this problem [4, 5].

Facing a series of challenges brought by the rapid change of knowledge content and the increasingly fierce international competition for highly qualified talents, it is an important and urgent issue for students to change their learning style and improve the quality of learning. In order to become qualified citizens in the 21st century, students must not only have a deep grasp of the subject matter, but also learn to learn and have better problem-solving, higher-order thinking, independent thinking and knowledge application skills. Therefore, in order to meet this requirement of social development, teachers of ideology and politics classes must change their teaching methods in the teaching pro-

cess, give full play to the main position of students in the learning process, provoke students to think and research deeply about the knowledge and content they have learned, and promote students' deep learning cite [6, 7]. The problem solving based on real situations is towards deep learning practice exploration, which is inseparable from the high quality classroom teaching problem design of ideology and politics class teachers. It can be said that good problems are favorable pushers to promote students' deep learning and play an important role [8].

On the one hand, the majority of English learners in China still use language repeaters, MP3 players, cell phones and other portable devices to assist in learning spoken English, but these methods are not able to perform However, none of these devices can recognize the pronunciation, and they are limited to the function of following the pronunciation, and they cannot directly give the learners reasonable and objective pronunciation evaluation and feedback guidance. Feedback and guidance cannot be given directly to the learners [9]. Moreover, due to the limitation of technology, some CALL systems at home and abroad mainly focus on the learning of words and grammar. The system has only one or two evaluation indexes as the basis for evaluation, which has certain functional defects and can only give learners an overall rating. This has certain functional shortcomings and can only give learners an overall rating [10, 11].

2. Speech Signal Pre-Processing

The purpose of these operations is to eliminate the effects on speech signal quality due to high harmonic distortion, high frequencies, and aliasing of the human vocal cords themselves and the speech signal acquisition equipment [12].

2.1. Pre-aggravation

The filter response function is shown in Eq (1):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{1}$$

The input speech signal $x(n)$ as:

$$y(n) = x(n) - \alpha x(n-1). \tag{2}$$

2.2. Split Frame

It can be regarded as a quasi-steady state process because the speech signal is relatively stable within a short-time range (generally 10 30ms), i.e., the speech signal has short time smoothness [13–15].

2.3. Add Window

To enhance the speech waveform around sample n and attenuate the rest of the waveform calculated by Eq (3):

$$Q_n = \sum_{m=-\infty}^{\infty} T[s(n)]\omega(n-m), \tag{3}$$

where T denotes some transformation, either linear or nonlinear, Q_n is a time series obtained after all segments are processed [16], rectangular and Hanning windows, which are defined as:

1. Hanming Window

$$\omega(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \leqslant n \leqslant N-1. \tag{4}$$

2. Rectangular window

$$\omega(n) = 1, 0 \leqslant n \leqslant N - 1. \tag{5}$$

3. Hanning Window

$$\omega(n) = 0.5\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right], 0 \leqslant n \leqslant N - 1. \tag{6}$$

The rectangular window has a narrow main flap with high frequency resolution, but also because it has a high side flap easily leads to more serious interference between adjacent harmonics, and sometimes superimposed and sometimes canceled within the adjacent harmonic interval to produce serious leakage [17].

The endpoint detection results for the sentence will be in the place where we always put it. Are shown in Figure 1. the results, it can be seen that the endpoint detection with the double threshold comparison method have better results.

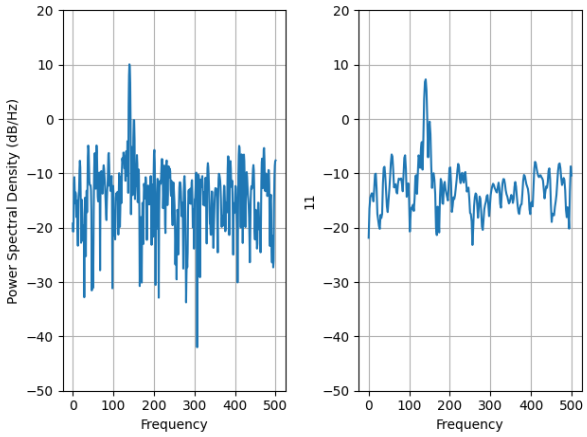


Figure 1. Sentence Will Be Placed Where We Often Put It

3. The Proposed Deep Learning Model

DBN is used to pre-train the weights of the generated model and then fine-tune the network with a back-propagation algorithm to obtain a better-performing network model. Numerous experiments have shown that internalizing the weight of a multilayer perceptron with a DBN of the appropriate configuration often gives much better results than random initialization [18]. DBN as shown in Figure 2. Essentially, the DBN is trained layer by layer on the RBM to obtain a global better initial parameter, thus improving the network performance. A large number of experiments also demonstrate that DBN can solve the problems of traditional BP networks: the need for a large set of training samples with markers, slow convergence, inappropriate parameter selection leading to the network falling into a local optimum.

Then the Wake-Sleep algorithm is used to adjust all the weight. The cognitive and the generative are made to agree as much as possible, i.e., the topmost representation of the generative is able to recover the bottom node as correctly as possible [19].

The training process of Deep Learning is as follows:

1. Unsupervised hierarchical training of each layer parameters using unsolicited data (calibrated data are also possible). The main difference from traditional neural networks is that this step is equivalent to the feature learning process [20].
2. Top-down supervised learning (that is, by going through the data with labels for training. The errors are transmitted top-down layer by layer to fine-tune the model parameters).

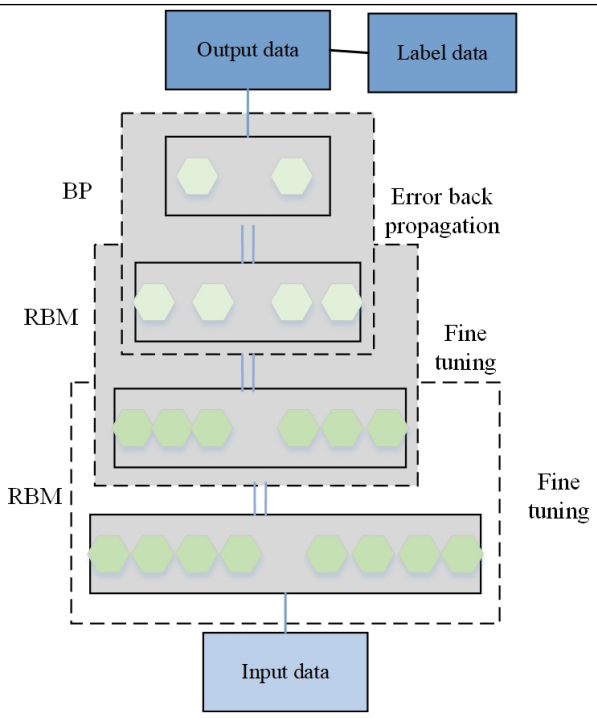


Figure 2. DBN Model

Supervised learning is used to further tune the parameters of the whole multilayer model. Unlike the random initialization process of traditional neural networks, the initialized parameters of DL are obtained by learning the structure of the input data in the first step and are not randomly initialized, thus the initial values are closer to the global optimum and thus better results can be achieved.

4. Experimental Simulation and Result Analysis

4.1. Data Source

This paper uses the Spoken Arabic Digit dataset from the UCI machine learning repository, constructed by the Automatic Signaling Laboratory at Badji-Mokhtar University. The dataset is the pronunciation of Arabic digits after extraction of 13-order MFCC feature parameters, and consists of a total of 8800 speech data (88 individuals pronouncing 10 Arabic digits, each repeated 10 times), pronounced by 44 men and 44 women between 16 and 40 years old.

Before MFCC feature parameter extraction, the parameters to be set are sampling rate 16KHz, 16-bit encoding, hamming window plus window function [21].

English sentence data source, the subjects of this paper was university students of our university, 24 in total, 15 male and 9 females. The subjects were recorded using the recording software CoolEdit, with a sampling rate of 16 KHz and 16-bit coding. A total of 10 sentences were recorded, all of which were commonly spoken English sentences

4.2. Speech Recognition Experiment

In order to verify the effectiveness of the model in this paper, we experimentally compare the recognition rate of this model with other models for non-person-specific solitary word recognition. In order to verify the effectiveness of this model, we compare the recognition rate of this model with other models for non-person-specific isolated word recognition. The Spoken ArabicDigit dataset from the UCI machine learning repository is used, which consists of 8800 Arabic digital speech data (88 individuals pronounce 10 Arabic digits, each digit is repeated 10 times). 6600 pronunciation from the first 66 individuals are used as the training set, and 2200 pronunciation from the last 22 individuals is

used as the test set [22].

For the same Spoken Arabic Digit dataset from the UCI machine learning library, [23] proposed a new Tree Distributions Approximation based on Graphical Tree Structure (TDA-GTS) and is similar to the Tree Distributions Approximation based on Maximum Weight Spanning Tree (TDA-MWST), the traditional Discrete Hidden Markov Model (DHMM), and the Continuous Continuous Markov Model (CMM). Model (DHMM), and a Continuous Density Hidden Markov Model (CDHMM) is compared and the recognition effect is improved. [24] proposed a K-means algorithm based on selective weights and Thresholds (KASWT), and compared it with BP_Adaboost algorithm, and the recognition effect was improved.

Here, the model in this paper is compared with the above models, and the comparison results of their recognition rates are shown in Table 1.

Model/Index	Recognition rate
DHMM	90.79%
CDHMM	94.09%
TCA-MWST	93.16%
BP_Adaboost	93.09%
KASWT	92.68%
Paper model	96.64%

Table 1. Comparison of Recognition Rates Under Different Models

As showed in Table 1, the recognition rate of the DBN model constructed in this paper is 96.64%, which is better than the above models. Therefore, the DBN-based speech recognition model established in this paper is reasonable and effective, and can be further used in speech pronunciation quality evaluation.

4.3. Speech Recognition Experiment

The purpose of the speech evaluation experiment is to verify the performance of the English speech pronunciation quality evaluation model and method proposed in this paper, the method is as follows:

$$A_{\text{Consistency rate}} = \frac{\text{Number of samples consistent between machine evaluation and manual evaluation}}{\text{Total number of samples}}. \tag{7}$$

The adjacent agreement rate is the ratio of the sum of the number of samples in which the machine evaluation and the manual evaluation agree and are adjacent to the total number of samples. Where "adjacent" is defined as the difference of one level between the ratings of machine and manual evaluations. The specific calculation method is as follows:

$$A_{\text{Adjacent consistency rate}} = \frac{\text{N} + \text{A}}{\text{Total number of samples}}. \tag{8}$$

4.4. Manual Evaluation

According to the pronunciation quality characteristics, we set four different levels for different evaluation indexes (pitch, speed, rhythm and intonation) and the overall evaluation situation.The detailed evaluation grades and corresponding evaluation standards are shown in Table 2.

The manual evaluation was done by two experienced university English teachers. They evaluated each of the 24 They evaluated each of the 10 recorded common English utterances of our university students, including 4 evaluation indicators of intonation, speed, rhythm and The four evaluation indexes and the overall evaluation were.

Grade	Intonation	Speed of speech	Rhythm	Intonation	Population
A	The content is complete and accurate	Moderate speed	Stress, accurate pronunciation and strong sense of rhythm	Accurate and natural intonation	Excellent pronunciation
B	The pronunciation is relatively clear and fluent	Speak a little faster (slower)	Stress pronunciation is more accurate and has a good sense of rhythm	The intonation is more accurate and natural	The pronunciation is generally good
C	There are pronunciation errors that affect understanding	Speak fast (slow)	The accent pronunciation is general and has a certain sense of rhythm	The intonation is basically accurate, but not natural enough	General grasp of pronunciation
D	There are serious pronunciation errors affecting understanding	Speak too fast (slow)	Accent pronunciation is wrong, the number of accents is too much (less), and the sense of rhythm is poor	The intonation is inaccurate and unnatural	Poor overall grasp of pronunciation

Table 2. Artificial Evaluation Level and Evaluation Criteria

Considering that the subjectivity of the teacher in the manual evaluation process may have an impact on the evaluation results, this paper uses the Pearson correlation coefficient to test the reliability of the manual evaluation results.

Further, the evaluation results of the two teachers were averaged (rounded) to obtain different students' different sentences for each evaluation index and the overall score as the final manual evaluation result.

4.5. Testing of the Evaluation Model

The regression analysis method uses mathematical and statistical methods to establish statistical models to study the statistical relationships between the variables of objective things , through a large number of tests and Through a large number of tests and observations of objective things, the statistical regularities hidden in those seemingly uncertain phenomena are searched for model predictions, as in Eq. (9).

$$\text{Score} = \text{AccuracyScore} \times 0.44 + \text{SpeedScore} \times 0.106 + \text{Rhythmscore} \times 0.341 - 0.397. \quad (9)$$

The F-test of the significance of multiple accuracy score, speedscore, rhythmscore and intonationscore are clear about the random variable score as a whole.Significant impact. According to the analysis of variance table calculated, indicating that the regression equation is significant, and the score has a significant linear relationship with accuracyscore, speedscore, rhythmscore and intonationscore, that is, the probability of error in the judgment that the four evaluation indexes have a significant linear impact on the total score is only 0.003 as shown in Table 3.

Index/Difference level	Consistency rate	Adjacent consistency rate	Pearson
Intonation	87.5%	100%	0.722

Table 3. Overall Evaluation Experimental Results

In the regression analysis method, the t-test was used to test the significance of the regression coefficients. The table of regression coefficients calculated by SPSS software shows that at the significance level, AccuracyScore, SpeedScore,RhythmScore, and IntonationScore passed the significance test indicating that all four evaluation indicators had a significant effect on the total score have a significant effect on the total score.

Further, Eq. (9) was used to evaluate a total of 240 sentences for 10 sentences for 24 students. The experimental results are shown in Figure 3, the distribution of different sample values is different. For example, the distribution of sample value 5 is 6, which is also what we expect. 210 samples have the same rating between machine, and no samples have the difference of two or three levels, the overall agreement rate between machine and manual evaluation is 87.5%, and the adjacent agreement rate is as high as 100%, which indicates that machine evaluation and manual evaluation have a strong correlation, indicating that machine evaluation has a strong correlation with manual evaluation.

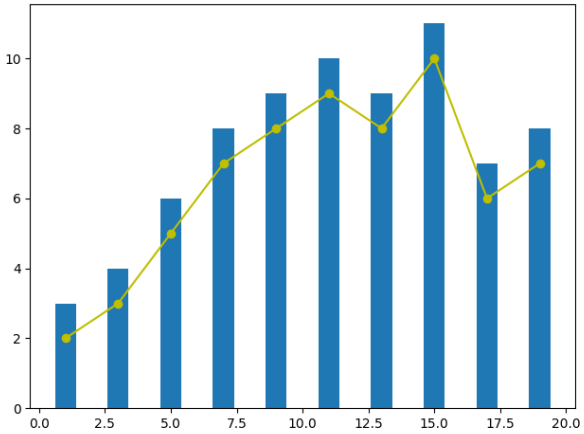


Figure 3. Graph of the Overall Evaluation Difference Between Machine and Human

5. Conclusions

This study advances the field of speech signal processing and evaluation through the integration of deep learning techniques and comprehensive evaluation methodologies. By leveraging DBN and rigorous evaluation criteria, we have developed a robust model for English speech pronunciation quality assessment. Our experiments demonstrate the superior performance of the proposed DBN-based model in speech recognition and evaluation tasks, outperforming existing methodologies. However, we acknowledge the need for further exploration to address limitations such as dataset size and evaluation methodology. Nevertheless, our findings lay a solid foundation for the continued refinement and application of deep learning approaches in speech processing and evaluation, with potential implications for language teaching and learning practices.

Conflict of Interest

Author declares no conflict of interests.

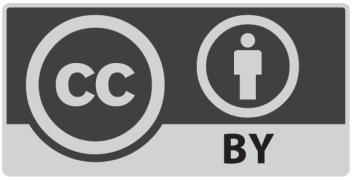
References

1. Wang, W., Meng, L., Wu, L. and Hu, J., 2020, March. Research on the main problems and countermeasures in the construction of university library informatization. In *IOP Conference Series: Materials Science and Engineering* (Vol. 799, No. 1, p. 012027). IOP Publishing.
2. Buigues, J.L.I., 2020. General Appraisal and Genesis of Regulatory Instruments in the Field of Civil and Commercial Law. In Coherence of scope of application: *Eu Private International Legal Instruments* (pp. 13-26). Schulthess Éditions Romandes.
3. Mu, L., Chen, W., Zhang, Y., Chen, X. and Li, W., 2022. A framework of analytical methods for horizontal behaviours of monopiles under VHM loads in sand. *Marine Georesources & Geotechnology, 40*(3), pp.349-360.
4. Lowe, A.A., Phan, H., Hall-Lipsy, E., O'Shaughnessy, S., Nash, B., Volerman, A. and Gerald, L.B., 2022. School stock inhaler statutes and regulations in the United States: a systematic review. *Journal of School Health, 92*(4), pp.396-405.
5. Ma, J., Ma, Y., Liu, Y., Zhai, G., Liu, S., Liu, H., Yue, G., Lan, X., Feng, Y., Qiu, X. and Zhang, P., 2022. Potassium Permanganate–Based Controlled Release Beads to Remediate Groundwater Pollution: Alkylbenzene Degradation and Permanganate Release Kinetics. *Water, Air, & Soil Pollution, 233*(8), p.323.

6. Vadiati, M., Rajabi Yami, Z., Eskandari, E., Nakhaei, M. and Kisi, O., 2022. Application of artificial intelligence models for prediction of groundwater level fluctuations: Case study (Tehran-Karaj alluvial aquifer). *Environmental Monitoring and Assessment, 194*(9), p.619.
7. Bradley, R.A., 2022. Assessing the effectiveness of several passive design strategies using the CIBSE overheating criteria: case study of an Earth Brick Shell House in Johannesburg, South Africa. *Architectural Science Review, 65*(3), pp.232-246.
8. Aiash, A. and Robusté, F., 2022. Traffic accident severity analysis in Barcelona using a binary probit and CHAID tree. *International Journal of Injury Control and Safety Promotion, 29*(2), pp.256-264.
9. Zeynolabedin, A., Olyaei, M.A. and Zahmatkesh, Z., 2022. Application of meteorological, hydrological and remote sensing data to develop a hybrid index for drought assessment. *Hydrological Sciences Journal, 67*(5), pp.703-724.
10. Bucking, S., Rostami, M., Reinhart, J. and St-Jacques, M., 2022. On modelling of resiliency events using building performance simulation: a multi-objective approach. *Journal of Building Performance Simulation, 15*(3), pp.307-322.
11. Bernardo, V., Costa, A.C., Candeias, P., Costa, A. and Catarino, J., 2022. Development of expeditious methods for seismic assessment of pre-code masonry buildings in Portugal. *Earthquake Engineering & Structural Dynamics, 51*(9), pp.2036-2054.
12. Afroogh, S., Esmalian, A., Mostafavi, A., Akbari, A., Rasoulkhani, K., Esmaeili, S. and Hajiramezanali, E., 2022. Tracing app technology: an ethical review in the COVID-19 era and directions for post-COVID-19. *Ethics and Information Technology, 24*(3), p.30.
13. Cui, X., Wang, Z. and Hou, C., 2015. Analysis and countermeasures to the problem of ultrasonic sensor receives the ultrasonic signal asymmetric. *Chinese Journal of Sensors and Actuators, 28*(1), p.2015.
14. Wang, J., 2020. Speech recognition of oral English teaching based on deep belief network. *International Journal of Emerging Technologies in Learning (Online), 15*(10), p.100.
15. Jin, S.C., Kim, D., Cho, S. and Sohn, S.B., 2018. Deep learning-based lip analysis system. *JP Journal of Heat and Mass Transfer, 15*(1), pp.29-33.
16. Zhang, Z., Zhang, C., Li, M. and Xie, T., 2020. Target positioning based on particle centroid drift in large-scale WSNs. *IEEE Access, 8,* pp.127709-127719.
17. Li, H., Zeng, D., Chen, L., Chen, Q., Wang, M. and Zhang, C., 2016. Immune multipath reliable transmission with fault tolerance in wireless sensor networks. In Bio-inspired Computing–Theories and Applications: 11th International Conference, BIC-TA 2016, Xi'an, China, October 28-30, 2016, Revised Selected Papers, Part II 11 (pp. 513-517). Springer Singapore.
18. An, P., Wang, Z. and Zhang, C., 2022. Ensemble unsupervised autoencoders and Gaussian mixture model for cyberattack detection. *Information Processing & Management, 59*(2), p.102844.
19. Qiao, L., 2022. Teaching design of online ideological and political course based on deep learning model evaluation. *Scientific Programming, 2022*, pp.1-8.
20. Cengiz, B.C., 2023. Computer-assisted pronunciation teaching: An analysis of empirical research. *Participatory Educational Research, 10*(3), pp.72-88.
21. Dai, Y. and Wu, Z., 2023. Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: a mixed-methods study. *Computer Assisted Language Learning, 36*(5-6), pp.861-884.
22. Smirkou, M., 2023. The Acquisition of English Suprasegmentals: A Computer-assisted Language Learning Approach1. Bridging Language Boundaries-Explorations in Communication across Borders, 15, p.105.

23. Sadiq, A.H.B., Razaq, H.R. and Mustafa, K., 2023. Tutoring Speech Organs with Computer-Assisted Language Learning. *Jahan-e-Tahqeeq, 6*(3), pp.95-107.

24. Kruk, M. and Pawlak, M., 2023. Using internet resources in the development of English pronunciation: the case of the past tense-ed ending. *Computer Assisted Language Learning, 36*(1-2), pp.205-237.



© 2024 the Author(s), licensee Combinatorial Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)