Study on the Effectiveness of the ASR-Based English Teaching Software in Helping College Students' Listening Learning

https://doi.org/10.3991/ijet.v12.i08.7142

Hongyan Zhao
Xi'an International University, Xi'an, Shaanxi, China
13759991618@163.com

Abstract—The development of automatic speech recognition (ASR) technology makes the human-computer interaction possible. Considering the deficiencies currently in the English pronunciation teaching, this paper employs relevant principles of speech recognition, proposes the design of an automatic English pronunciation grading system in the English teaching field, describes in detail the system structure, functions and processes and introduces the key technologies and steps to implement the system: dynamic time warping algorithm, establishment of corpuses, consonant/vowel (C/V) segmentation technology and grading standards. Results of a small-scale test show that the system is useful to testing the English pronunciation of international students.

Keywords—speech recognition, English learning, automatic grading, dynamic time warping, phoneme segmentation

1 Introduction

Language is a main tool in interpersonal communication. Under the context increasing international economic and trade activities, more and more people begin to learn the languages of other countries, the most popular of which is undoubtedly English. At present, English teaching is becoming a hot spot in the education field. However, the traditional English teaching approach is not very effective and does not meet the requirements of good English education. In this case, a computer-aided language learning (CALL) technology has been developed and widely used in English education and learning [1]. However, in the application of this technology, due the lack of experience, many teachers put too much emphasis on English reading and writing and pay little attention to speaking. As a result, students are often learning "dumb English". At present, people all over the world are actively learning English, and the market demand for English learning is also growing, but English teacher resources are few and cannot meet the teaching needs. In language learning, pronunciation training is an important part. Through this training, students can find their own pronunciation errors so that they can correct them in a timely manner [2].

During classroom teaching, teachers can promptly find students' pronunciation errors and correct them in time. However, influenced by the traditional teaching philosophy, many English teachers have a certain "fault tolerance" of students' poor pronunciation. As a result, many students do not pronounce correctly. Although these international students can have simple dialogues with teachers and the communication is effective, they will find it very difficult to communicate smoothly with common foreigners [3-5]. Therefore, in English pronunciation teaching, there should be a scientific and reasonable English pronunciation evaluation system to evaluate students' pronunciation.

2 ASR Technology

This technology mainly uses the pattern recognition principle to recognize speech, and its theoretical framework is shown in Fig.1. As can be seen from Fig. 1, this technology mainly consists of several units such as speech signal preprocessing, feature extraction and modeling. Below is a detailed analysis.

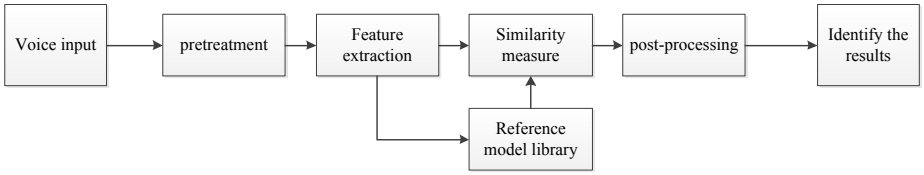

Fig. 1. Speech Recognition Principle

The training using ASR system can be divided into two stages - training and recognition. Both of these stages require preprocessing the original voice input and extracting the corresponding features. The preprocessing module mainly processes the original speech signals and filters out some noise and background hiss and at the same time analyzes the sub-frames of these voice signals and reprocesses them [6-8].The feature extraction module mainly allocates and adjusts the corresponding acoustic parameters, and calculates the voice features in order to extract the key feature parameters of signal features for subsequent speech processing [9]. In speech processing, this system mainly uses parameters like the amplitude, energy, zero-crossing rate, cepstrum coefficient and short-term spectrum. By rationally adjusting these parameters, the system will achieve speech recognition. Thus, the choice of features is the core of the system construction.

In the training stage, the user first inputs a certain amount of training speech. The system performs preprocessing and feature extraction, and determines the required feature vector parameters, and then establishes the corresponding speech reference model library based on these parameters. If the conditions are not sufficient to establish one, the system can also choose the existing model library for appropriate corrections. In the recognition stage, the system compares the similarity measures between the input feature vector parameters and the template in the library to determine the types with the highest similarity, and outputs them as the results. Later, the system

needs to further process the recognition result candidates, and under the relevant semantic and phonetic constraints, it determines the final recognition result [10-12].

3 Automatic English Pronunciation Grading System

Considering the language-learning pattern, the pronunciation learning using this model should be from easy to difficult, and the learning process should be divided into the following specific stages: the first step is to learn the basic pronunciation, mainly including English vowels and consonants; the second step is to learn the pronunciation of words and combinations of vowels and consonants and the tone and intonation of sentences; and the third step is mainly to train the pronunciation of phrases and sentences, and teach students to master co-articulation and rhythm.

Specifically, this system can be divided into four modules, which are highly independent of each other. Details of the modules and relevant functions are as follows: ① standard model training module: it is mainly used to establish a standard corpus. This system can compare the learner's pronunciation against the pronunciation in the standard corpus to find out the similarity and give a corresponding score based on the comparison results. Therefore, for this system, the training and establishment of a standard model is one part of the main work. During the establishment, usually standard phonetic materials will be needed to train the standard corpus in order to ensure the scientific validity of the results [13]. ② Expert grading module: it is mainly designed to evaluate the pronunciation grade of the learner. The score given out by this system may not be directly comprehensible, as it cannot measure the actual pronunciation level of the learner in a convenient manner. Therefore, it is necessary to establish grades for pronunciations in the non-standard corpus against the standard pronunciations, and evaluate the pronunciation based on these grades to make the results easier to understand. ③ Automatic grading module: it is mainly designed to compare and analyze the pronunciations in the standard corpus and those of the learner. Based on speech recognition, it employs some similarity algorithm to determine the indicators of certain pronunciations. This module collects the learner's pronunciation, preprocesses it and determines its features, and then segment the voice into a number of elements according to the established element model. After that, it grades the learner's pronunciation based on the similarity algorithm according to English pronunciation features [14-16]. ④ Error analysis module: it is mainly designed to determine the score of the learner's pronunciation based on the expert knowledge base, and give advice on how to correct some of the pronunciation errors. Students can use this module to correct their own pronunciation problems and get some good advice and help. However, this process requires a lot of expertise, so it is still under improvement. The flow chart schema of this system is shown in Fig.2.

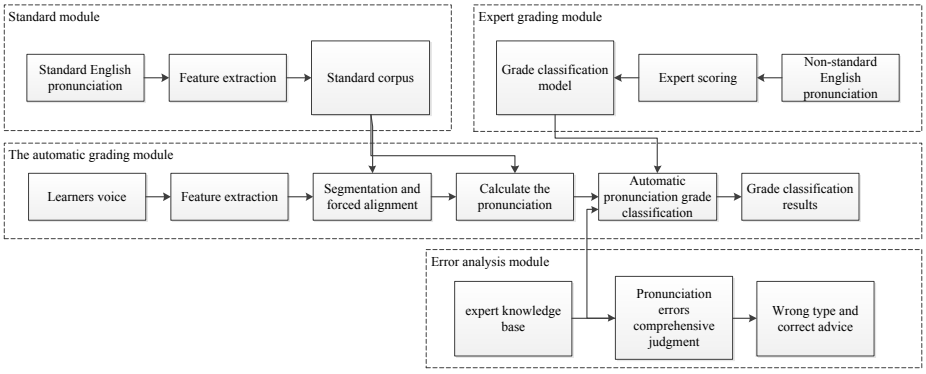

Fig. 2. English Automatic Grading System and Process Flow Chart

4 Key Technologies to Implement the System

4.1 DTW (dynamic time warping) algorithm

In this recognition process, if the reference template and the input template are directly compared without pre-processing, the result will not be very effective. This is mainly because there is certain randomness with the speech signals, and different people make different pronunciations, and even the same person may pronounce differently at different moments with different time lengths [17-20]. To this end, it is necessary to carry out time correction processing, that is, using the dynamic programming technology to convert the global optimization problems into simple local optimization problems, and then solving the problems one by one. Below is a detailed discussion of this process.

For convenience of analysis, the feature vector sequence of the reference template is expressed as $X= \{X_1, X_2, \ldots, X_I\}$, and the corresponding input speech feature vector sequence is $Y = \{Y_1, Y_2, ..., Y_J\}$. Let the time warping function be $C= \{C(N)\}$, where N is the length of the corresponding path; the distance between the two $d(X_i(n), Y_j(n))$ is the local matching distance. When this DTW algorithm is used in the calculation, the sum of the weighted distances can be minimized based on local optimization, and the specific formula is as follows:

$$D = \min_{C} \frac{\sum_{n=1}^{N} [d(x_{i(n)}, y_{j(n)}) \cdot W_n]}{\sum_{n=1}^{N} Wn} \tag{1}$$

When we use the DTW algorithm in speech computation, we need to first determine the optimal time warping function, and use this function to accurately map the timeline of the speech to be tested to the reference model to minimize the accumulated distortion. This algorithm was initially used in the pattern recognition field, and later widely applied in the English speech recognition field. To date, it has become a

classic algorithm, which can efficiently and quickly recognize some isolated words without doing much computation. Therefore, it has been extensively applied in the speech recognition field.

4.2 Writing a new document with this template

When building the English pronunciation unit model and the corresponding grading model, we need corpuses for training. During training, we can choose one of the corpuses as needed: the two corpuses are standard and non-standard pronunciation corpuses respectively. The former is mainly used to train pronunciation units. To build this corpus, we choose 10 foreign teachers and Chinese international students whose pronunciations are standard to do the pronunciation unit training so that the pronunciations provided by this corpus will be not only standard, but also easier for students to accept. Based on practical experience, if a corpus only contains the pronunciations of native speakers, international students will not get high scores, and thus they will not be much interested in English learning. Therefore, an appropriate model needs to be built based on actual conditions to achieve better matching effect.

In building the standard corpus, we choose the robust model training approach, that is, to compare the pronunciations of a word by multiple persons, average the corresponding vectors along the DTW path and obtain the final model. The establishment and training process of the corpus is as follows:

For a specific word, let X_1 and X_2 represent the feature vector sequences of the first and the second pronunciations respectively, and determine the distortion score of the two $d(X_1, X_2)$ through the DTW algorithm. If the results are smaller than the corresponding threshold value, it can be deemed that the pronunciation feature vectors of the two meet the consistency requirement. Then determine the time-warped average of the two and obtain a new model Y. All elements y_k can be calculated by the following method:

If Ty is the optimal path length in the DTW algorithm, the corresponding sequence is expressed as follows:

$$(i(1), j(1)), (i(2), j(2)), \ldots\ldots, (i_{Ty}(Ty), j_{Ty}(Ty)) \tag{2}$$

Components of the new model can be determined through the following formula:

$$y_k = \frac{1}{2}(x_{1i(k)} + x_{2j(k)}), k = 1, 2, \ldots\ldots, \mathrm{Ty} \tag{3}$$

The non-standard corpus is designed to help build an appropriate manual grading model and provide certain reference. In this paper, in order to build a non-standard grading model, we choose the speeches of 50 foreign international students and foreign teachers with significantly different English pronunciation levels. Then teachers sort out and grade these pronunciations.

4.3 C/V segmentation algorithm

In C/V segmentation, the English speech evaluation system needs to be used. This system can normalize the speeches of the learner and in the standard corpus, obtain consistent speech feature parameters, then does matching computation of the similarities between the two on a C/V segmented basis, and at last determines the Euclidean distance between the two to reflect the pronunciation level of the learner. However, for convenience of analysis, usually the system does global matching computation of the two speeches and obtains the result based on C/V ratio. In an English syllable, the proportions of vowel and consonant are quite different, and the proportion of consonant is far smaller than that of vowel. As a result, some light and short consonants are very likely to be "drowned" by those stressed and long consonants. In this case, the consonant feature of a syllable will be lost, causing an incorrect matching result. To solve this problem, the system has to separate the vowels and consonants first, then does matching computation of the two separated parts respectively, and at last determine the final matching result on a weighted basis [21-22].

The allocation principle in this algorithm is as follows: within one phoneme, the feature vectors of speech frames are highly similar, but in different phonemes, the feature vectors of speech frames are significantly different. We can use this characteristic to distinguish vowels and consonants. Meanwhile, we can measure the speech differences by the distance between phonetic segments.

For convenience of analysis, we usually call the overall difference between two speech segments the distance between the two, i.e. distance between segments. The calculation of this distance is to measure the similarity. This type of calculation usually involves time frames. Based on the distance between segments, we can calculate the phonetic segment consisting of a certain number of continuous speech frames, and use the distance between segments to reflect the feature difference between speeches.

Let a and b be the two segments in one continuous speech. The number of time frames in these two segments are N_a and N_b, and there can be overlaps between the two. The feature vectors of time frames in these two segments are X_i and Y_i respectively, and the Euclidean distance between the two is d_{ij}. Then we can calculate the distance between Segment a and b using the following formula:

$$D = \frac{1}{N_a N_b} \sum_{i=1}^{N_a} \sum_{j=1}^{N_b} d_{ij} \tag{4}$$

Based on the distance between segments, we can determine the difference between the two speech segments. If the adjacent two speech segments are one phoneme, the distance between the two will be short; otherwise, the distance will be long. In this way, we can separate the vowel and the consonant in a syllable based on the distance between segments.

4.4 Grading standards

The English pronunciation grading system plays an important role in English learning. It can determine the learner's pronunciation level and give appropriate tips. To grade their pronunciation, we can take standard pronunciation as the basis, calculate the distance between the standard pronunciation and the learner's and give a pronunciation score based on this. This approach is convenient, but there are still some problems – the score is assertive and unstable, and the result is related with the learner's personal condition, difficult to understand and inconsistent with people's perception. Therefore, usually the expert grading method is used, which converts the measured pronunciation score to expert's score on the basis of mapping, and gives the final pronunciation result on a comparative basis. This method needs to classify the pronunciation quality into grades like "very good", "good", "average" and "poor" based on experts' score using expertise. This method has such merits like clear meaning, conformity to perception habits and stable results, but the results can also be subjective. When this method is used to evaluate the pronunciation, the user should choose articulation and naturalness, etc. as the evaluation indicators.

5 Conclusions

This paper mainly studies the application of speech recognition software in English listening learning. In the research, in order to verify the consistency between the scores given by the learning software and experts, we selected 20 Chinese international students at varying English levels as the research objects and tested their pronunciations. We chose English junior- and intermediate-level textbooks as the testing materials and selected over a thousand words covering English syllables and tones from these materials. We tested the students in a language lab. Each test object read 20 random words, which were recorded and graded by the system. At the same time, teachers graded their pronunciations. We compared the results obtained from the two grading methods, and found that this system can accurately evaluate the English pronunciations of international students at different levels, and that the results are accurate and consistent with the teachers' scores, thus this system is helpful in training English pronunciations.

The test results also reveal some problems in this system that need improvement. For example, the system is not stable enough – it lacks the standard training speech data in an environment with interference, thus the score for some rarely used words is often zero; judging from actual practice, the grading mechanism used in this system is still not scientific enough and the scores are not clear enough, and thus the results are not stable. In future research, we need to work on the consistency and stability of scoring, establish a more standardized expert system and at the same time improve its functions so that the system can correct learners' mistakes while grading their pronunciations. In this way, the applicability of this system will be improved, which will lay a good foundation for the promotion of the system in the future.

6 References

[1] Li C. (2012). Multimedia computer assisted instruction in college English teaching. Journal of Hunan University of Science & Engineering, 2006, 754-757.

[2] Shi B. (2009). Empirical research on feasibility and effect of computer-aided College English teaching model. International Conference on Information Engineering & Computer Science, 1-4, 236-241. https://doi.org/10.1109/ICIECS.2009.5362868

[3] Wang Y., Bao Y. (2010). Research on human-computer interaction of English teaching at local area network. International Conference on Computer Application & System Modeling, 2, 718-722. https://doi.org/10.1109/ICCASM.2010.5620224

[4] Yang Q.H. (2013). Research of college English teaching based on computer network technology. Informatics and Management Science III, 206, 375-382. https://doi.org/10.1007/978-1-4471-4790-9_48

[5] Wu X., Chen H. (2011). Computer-based educational game and its application in college English teaching in China. International Conference on Multimedia Technology, 3008-3011. https://doi.org/10.1109/ICMT.2011.6001973

[6] Chen X.H. (2014). The research on English autonomous learning monitoring theory and application in the network environment. Applied Mechanics and Materials, 644-650, 6079-6082. https://doi.org/10.4028/www.scientific.net/AMM.644-650.6079

[7] Song J., Wu H. (2012). The English teaching model of cooperative learning in the network environment in higher vocational education. Advances in Computer Science Environment Ecoinformatics & Education, 218, 100-104, August 2012. https://doi.org/10.1007/978-3-642-23357-9_19

[8] Sun Q.L. (2015). Information under the network environment using computer information security technology. International Conference on Intelligent Transportation, 474-477. https://doi.org/10.1109/ICITBS.2015.122

[9] Xu M.Q. (2011). The research on out-of-class autonomous English learning in computer- and network-assisted environment. Advances in Intelligent and Soft Computing, 108, 453-459. https://doi.org/10.1007/978-3-642-24775-0_71

[10] Yang D.L., Zheng H. (2010). Research on the framework of new college English teaching mode integrating cooperative and autonomous learning in the network multimedia environment. ICETC 2010-2010 2nd International Conference on Education Technology and Computer, 3, 256-259. https://doi.org/10.1109/ICETC.2010.5529552

[11] Chen X.H. (2014). The research on English autonomous learning monitoring theory and application in the network environment. Applied Mechanics and Materials, 644-650, 6079-6082. https://doi.org/10.4028/www.scientific.net/AMM.644-650.6079

[12] Han S.F., Miao S. (2011). On college English teaching of writing in the network environment. 2011 International Conference on Multimedia Technology, 588-590. https://doi.org/10.1109/ICMT.2011.6002176

[13] Li S. (2011). Survey research on college students' English learning anxiety in the computer network environment. ICCSE 2011-6th International Conference on Computer Science and Education, 1010-1012. https://doi.org/10.1109/ICCSE.2011.6028807

[14] Yang D.L., Zheng, H. (2010). Research on the framework of new college English teaching mode integrating cooperative and autonomous learning in the network multimedia environment. ICETC 2010-2010 2nd International Conference on Education Technology and Computer, 3, 256-259. https://doi.org/10.1109/ICETC.2010.5529552

[15] Zhou J., Chen, X.H. (2014). Analysis and simulation of computer virus propagation models in the network environment. Advanced Materials Research, 204-210, 433-436. https://doi.org/10.4028/www.scientific.net/AMR.204-210.433

[16] Wang Y., Bao Y. (2010). Research on human-computer interaction of English teaching at local area network. ICCASM 2010 - 2010 International Conference on Computer Application and System Modeling, 2, 718-722. https://doi.org/10.1109/ICCASM.2010.5620224
[17] Lei Z., Wei X. (2011). The analysis of English network computer aid test system model and technology selection. 2011 International Conference on Internet Technology and Applications, 1-4, 236-248. https://doi.org/10.1109/ITAP.2011.6006325
[18] Wang X.M., Yang Y.J., Wen X. (2009). Research and design of computer-aided English textbook evaluation system. Proceedings of the 1st International Workshop on Education Technology and Computer Science, 3, 913-917. https://doi.org/10.1109/ETCS.2009.741
[19] Rahkila M., Karjalainen M. (1999). Evaluation of learning in computer based education using log systems. Proceedings-Frontiers in Education Conference, 1, 16-21. https://doi.org/10.1109/FIE.1999.839266
[20] Yamamoto H., Ohtani A., Kado T. (1993). Development and evaluation of computer-mediated education systems for customer engineers. IFIP Transactions A: Computer Science and Technology, 35, 205-214.
[21] Bai Y. (2014). Comprehensive evaluation of education mode of university English model and swat analysis. Computer Modelling and New Technologies, 18(11), 1085-1088.
[22] Jia J., Chen W.C. (2009). The further development of CSIEC project driven by application and evaluation in English education. British Journal of Educational Technology, 40(5), 901-918. https://doi.org/10.1111/j.1467-8535.2008.00881.x

7 Author

Hongyan Zhao, postgraduate and associate professor, is an English teacher at Xi'an International University, 408 Zhangba Road, Shaanxi Xi'an, China. She has been engaged in English teaching for almost ten years and has taught more than ten courses such as English writing, advanced English, marketing English, business English, college English, etc. Her major research orientations are business English and English language teaching.

Article submitted 05 May 2017. Published as resubmitted by the author 13 June 2017.