Education Research Frontier
September 2020, Volume 10 Issue 3, PP. 184-188

English Speech Recognition and Multi-dimensional Pronunciation Evaluation

Jinwei Dong, Shaohui Li†

School of English Education, Guangdong University of Foreign Studies, Guangzhou 510006, China

†Email: shevyyy@163.com

Abstract

As one of the four major skills in English learning, speaking is an important part in English learning which getting more and more attention. The study set out to improve the accuracy of speech recognition by applying speech recognition technology to the evaluation of English pronunciation. On this basis, the study attempts to improve computer speech evaluation methods. This model employs indexes such as intonation, speech rate, rhythm and intonation, using multi-dimensional indicators as evaluation criteria to establish a more comprehensive and objective English speech evaluation model. English learners can use the evaluation systems to guide their self-study, and the system can have a profound influence on the practice of teaching spoken English.

Keywords: Spoken English; Speech Recognition; Pronunciation Evaluation

1 INTRODUCTION

In the teaching of English, the improvement of students' spoken English is getting more and more attention. Cultivating English listening and speaking abilities is one of the most important goals in developing students' English literacy. In the learning process of spoken English, real-time feedback on pronunciation is very necessary. Real-time feedback allows students to discover pronunciation problems in time and make corrections. Comprehensive pronunciation feedback can enable students to improve their spoken English with focuses and complement their shortcomings in pronunciation. This is not only beneficial to students' learning of spoken English, but also has positive influences on the practice of spoken English teaching and the testing and evaluation of it.

2 THE IMPORTANCE OF LEARNING SPOKEN ENGLISH AND THE PROBLEMS IN THE CURRENT TEACHING

As one of the four major skills, spoken English is an important part of English learning, and it is getting more and more attention. In the National Language Standard released in 2018, English language ability is defined as the ability to interpret and express intended meanings. Language activity is defined as what learners and users of English do when they communicate information through the use of language in a certain situation [1]. Therefore, learners need to use the knowledge of English to express and communicate in language activities on the basis of mastering grammatical rules in order to complete communicative activities successfully. Moreover, according to the input and output theory of second language acquisition, comprehensible language input is vital in the acquisition process. But it is still not enough to improve the learner's English abilities comprehensively. For learners, language output is as important as language input. And oral expression is one of the main ways of language output. In addition, the phenomenon of knowing English well but cannot use it still exists, which reflects the shortcomings in English education [2]. Timely pronunciation Evaluation is of great importance to improve students' spoken English. In order to help learners to know the quality of their pronunciation comprehensively, it is necessary to evaluate the accuracy of speech objectively and effectively. In the current spoken English teaching and learning practice, there are still many problems in evaluating the pronunciation of students.

First of all, in College English classes, it is difficult to realize the timely evaluation of the pronunciation of each student. The major teaching tasks in College English classes is not just the teaching of spoken English, which also

includes the learning of professional vocabulary, grammar, cultural background and so on. The teaching of spoken English and the correction of students' pronunciation are just parts of teaching. When students read articles aloud or answer questions, although teachers can point out and correct the problems of their pronunciation or intonation, it is unrealistic for teachers to correct each student's pronunciation during limited time of the lesson. In the College English class, the number of students is often large, and each student's level of spoken English and problems of pronunciation are often different. It is difficult for the professors to solve the pronunciation problems of each student in a well targeted way in classes. Furthermore, in terms of spoken English learning platforms, it is difficult for students to get comprehensive pronunciation evaluations. Limited by technology, most domestic spoken English learning platforms are mainly focus on the correction of words and the completeness of sentences retelling, and only one or two dimensions are adopted when evaluating pronunciation. In terms of the presenting of evaluation result, it is relatively simple. The evaluation results are often presented in the form of scores, and students often cannot figure out their specific pronunciation problems and make corrections based on the evaluation results. In addition, in the evaluation of spoken English, human rating is often adopted currently. Although human rating has its advantages, for example, raters can make detailed and professional judgments on the pronunciation of students based on their background knowledge, which is more flexible than machine scoring. But it also has many disadvantages. The subjectivity of human rating is relatively strong, because each rater is an individual with different experience and knowledge background. Despite the establishment of scoring standards, it is still difficult to ensure the objectivity of scoring. Even the evaluation is performed by the same rater, factors such as working hours, emotional state can also lead to deviations in the score.

In order to solve the above problems, the study set out to improve the accuracy of speech recognition by applying speech recognition technology to the evaluation of English pronunciation. On this basis, the study attempts to improve computer speech evaluation methods. This model employs indexes such as intonation, speech rate, rhythm and intonation, using multi-dimensional indicators as evaluation criteria to establish a more comprehensive and objective English speech evaluation model.

3 RESEARCH AT HOME AND ABROAD

Since the 1950s, with the vigorous development of computer technology, automatic speech recognition technology has gradually been applied to various fields, especially in the field of language learning. Researchers at home and abroad have made significant contributions to the study of speech recognition and pronunciation assessment. In terms of speech recognition, in the 1950s, foreign researchers began to study speech recognition. Its iconic beginning was the establishment of isolated speech recognition systems for specific persons, and its founder was Bell Labs. [3] By the 1960s, the dynamic time warping (DTW) algorithm was successfully used in the analysis of speech signals, which was a major breakthrough in the application of the algorithm in analyzing speech signals. By the 1970s, the technology of extracting characteristic parameters had gained significant progress. Subsequently, Baum et al. established a hidden Markov model by combining introduction and statistical methods. In the application of this model, Baker et al. of MIT further applied it to speech recognition. In the 1980s, neural network technology developed rapidly and was used in speech recognition. The application of artificial neural network is conducive to processing speech signals, which is of great significance.

In recent years, the deep learning method in the field of education have been adopted to the field of computer, and further applied to the research of speech recognition. Among the large body of research, the research of Microsoft Research Asia has contributed a lot to this field. By using the DNN-HMM hybrid network model to train the acoustic model, the goal of identifying a large number of words was achieved. The research on speech recognition in our country began in the 1960s by the Chinese Academy of Sciences. After the 1980s, the Chinese Academy of Sciences, Tsinghua University, Peking University and other institutions all established speech recognition research laboratories. In terms of pronunciation evaluation, Liang Weiqian and others from Tsinghua University proposed the PASS algorithm based on the HMM model at an early stage [4]. Using this algorithm to evaluate the quality of English pronunciation achieved considerable results. Regarding the research of factor-based evaluation algorithms, there was the improved algorithm proposed by Liu Qingsheng et al [5]. This algorithm improved the efficiency of the operation and made the computer score similar to that of professional raters. In terms of large-scale speech

evaluation, Li Mengtao et al. studied the scoring of question-answer in oral tests based on machines, which explored the feasibility and application of machine scoring [6].

With the advancement of pronunciation evaluation technology, more research in this field have appeared, including research of some automatic evaluation systems and spoken English test systems based on computers. Among them, there was the Changyan English learning system developed by Xunfei [6]. The above relevant research and applications indicate that relevant techniques of pronunciation evaluation are relatively mature. It can be applied to multiple scenarios such as autonomous learning, classroom teaching and evaluation.

4 MULTIDIMENSIONAL EVALUATION OF THE QUALITY OF ENGLISH SPEECH BASED ON SPEECH RECOGNITION TECHNOLOGY

4.1 Speech Recognition Technology

Generally speaking, speech recognition technology refers to semantic recognition and voiceprint recognition; in a narrow sense, it refers to the recognition of speech semantics, also known as automatic speech recognition. The preprocessing of speech signals is critical in speech recognition. The steps of speech signal preprocessing are speech signal digitization, endpoint detection, framing, windowing, and pre-emphasis. Preprocessing the speech signal is to extract the signal characteristics of the speech from the original speech of the speaker. Speech feature parameters are also necessary to evaluate the quality of English pronunciation, because the characteristics of speech need to be expressed by speech feature parameters. The comparison of speech is achieved by comparing the parameters of speech characteristics. To describe the characteristics of the speech signal, there are multiple parameters that can be adopted, such as linear prediction coefficient (LPCC) and Meier cepstrum coefficient (MFCC). Among these parameters, the Meier Cepstral Coefficient (MFCC) has the strongest noise resistance and stability. When evaluating speech, by comparing the characteristic parameters of the two, the similarities between the learner's pronunciation and the reference pronunciation can be obtained. However, a direct comparison between the two ignores the differences in the dimensions of speech length and speech rate, and the conclusions drawn are not objective. In order to improve the accuracy of comparison, it is necessary to use a certain discriminant algorithm to match the characteristic parameters of the two patterns. In speech recognition, dynamic time warping algorithm (DTW), algorithm based on hidden Markov model (HMM), artificial neural network (ANN) and deep neural network (DNN) are available. Among them, the application of deep neural network construction mode has unique and prominent advantages compared with other modes. First of all, the traditional neuron analog information processing system has a single level of division, while the speech recognition and evaluation system based on a deep neural network discriminant algorithm has a richer level of division. Secondly, the traditional neuron information processing system is too extensive, and the speech recognition and evaluation system based on deep learning is more targeted. Therefore, this study uses the discriminant algorithm of deep neural network. High-precision and high-speed speech recognition can provide a solid foundation for the next stage of pronunciation evaluation.

4.2 Pronunciation Evaluation

Speech evaluation is a branch of speech recognition. Speech evaluation is the analysis of speech to obtain an objective evaluation. In the 21st century, the setting of evaluation indicators for the quality of English pronunciation has been basically unified and standardized, and it is recognized worldwide. This evaluation system mainly combines multi-parameter pronunciation evaluation indexes such as rhythm, speed, rhythm, intonation and intonation of English pronunciation, and attempts to improve the comprehensive evaluation algorithm of pronunciation quality.

4.2.1 Rhythm Evaluation

From the perspective of linguistics, it refers to the phonological organization above the syllable level, which organizes various linguistic units into discourse or related blocks in discourse. Prosody features, also known as super segment features, mainly refer to speech rate, rhythm and intonation and so on. These features are composed of dynamic patterns of related quantitative parameters such as pitch, intensity and duration. Prosody can convey many

non-linguistic information of the speaker, including emphasis, pause, stress, and phrase combination. The non-linguistic information can often express the speaker's emotions, attitudes, tone and so on. It is also an important index that cannot be ignored in evaluating the quality of English pronunciation. Rhythm evaluation includes pace evaluation, speech rate evaluation, and intonation evaluation. Among them, due to the differences in the characteristics of Chinese and English, the pace of pronunciation is also very different, and the control of pace has become a major difficulty in speaking. Stress, division of sense-group will affect the pronunciation pace. Therefore, pace evaluation is an important indicator of English pronunciation quality evaluation. The evaluation of speech rate is determined by the length of speech. Intonation refers to the tone of speech, including the change of the tone and the change of humility. In this study, to achieve multi-dimensional evaluation and to provide feedback of English pronunciation, speech rate evaluation based on speech duration, rhythm evaluation based on short-term energy and paired variation index, and intonation evaluation based on fundamental frequency are adopted.

4.2.2 Pitch Evaluation

The standards of investigating intonation are accuracy, fluency and completeness of pronunciation. The accuracy of content is judged by means of MFCC coefficients, and a deep belief network model is established to recognize pronunciation. Furthermore, by calculating the correlation coefficient of the MFCC features of the standard sentence and the input sentence, the judgment of the fluency of pronunciation is realized. Finally, by synthesizing the evaluation results of the two, an overall evaluation of pronunciation quality is obtained. Based on the full use of speech recognition technology, combined with standardized pronunciation quality evaluation indicators, it constitutes an objective and efficient English speech evaluation system for multi-dimensional evaluation. This will help to improve the effect of the teaching of spoken English.

CONCLUSION

In recent years, the further development of science and technology has promoted the reform of oral English teaching and assessment methods. The English pronunciation evaluation system can be applied to multiple scenarios, and it can have a strong impetus to learners and teachers in English autonomous learning, classroom teaching, oral practice, language testing and evaluation. English learners can use the evaluation system to guide the self-study of spoken English, and the scientific and objective evaluation based on multiple dimensions can help learners improve their pronunciation with focuses. With the continuous development of informationization, English speech recognition and multi-dimensional pronunciation evaluation system will show its unique advantages in application. And it will be welcomed by more learners and English teachers.

ACKNOWLEDGEMENT

This study was financially supported by the Undergraduate Innovation Training Project of Guangdong University of Foreign Studies in 2020.

REFERENCES

[1] Wang Shouren. The Application of China's Standards of English Language Ability to College English Teaching[J]. Foreign Language Education,2018,39(04):1-4.

[2] Xu Jinfang. How to Say Goodbye to Dumb English in Junior Middle School English Teaching[J]. China After School Education,2019(33):118.

[3] L R Rabiner, B H Juang. Fundamentals of Speech Recognition[M]. New Fersey: Prentiee Hall,1993.

[4] Liang Weiqian, Wang Guoliang, Liu Jia, and Liu Runsheng. Phone-based Pronunciation Quality Assessment Algorithm[J]. Journal of Tsinghua University (Science and Technology),2005(01):5-8.

[5] Liu Qingsheng and Yu Dong. Research on a Speech Endpoint Detection Method[J]. Computer Engineering,2003(03):120-121+138.

[6] Li Mengtao, Yang Xiaoguo, Feng Guodong, Wu Min, Chen Jiliang, and Hu Guoping. Feasibility Study and Practice of Large-scale College English Speaking Test Reading Aloud Question Marking Machine[J]. Foreign Language World,2008(04):88-95.

AUTHORS

[1]**Jinwei Dong**, China, January 1966, graduated from the research center of foreign literature and culture of Guangdong University of Foreign Studies with a doctor's degree in literature in 2016. Professor of English, tutor of master's degree, Dean of College of English education. The main research fields are College English education, teacher professional development, legal linguistics.

[2]**Shaohui Li**, China, Feburary 12, 1998. The Undergraduate majoring in Education of Guangdong University of Foreign Studies in Guangzhou, China. The main study field is Computer Aided Oral English Evaluation.