







Education of Recognition Training Combined with Hidden Markov Model to Explore English Speaking

Hui Li¹ , Xinyu Zhang² , Ran Cui³ , and Na Wang⁴ 

¹Cangzhou Normal University, Cangzhou, Hebei 061000, China, lihuicangzhou@163.com

²Cangzhou Normal University, Cangzhou, Hebei 061000, China, flw@caztc.edu.cn

³Cangzhou Normal University, Cangzhou, Hebei 061000, China, zhangxin1408@163.com

⁴Cangzhou Normal University, Cangzhou, Hebei 061000, China, wangnacangzhou@163.com

Corresponding author: Xinyu Zhang, flw@caztc.edu.cn

Abstract. In order to study a better speech scoring mechanism to achieve a more accurate evaluation of the practitioner's spoken pronunciation, and better help the practitioner to find out the lack of oral pronunciation, this study studied HMM-based speaking training system. The system evaluates the learner's spoken pronunciation from the vocal segment, the super-sound segment and the perception domain of the speech signal and improves the correlation between the computer score and the expert manual score. Aiming at the difficulty of oral English teaching in current English teaching, this paper designed a spoken language training system based on automatic speech evaluation using advanced computer technology, network technology, natural language processing technology, speech processing technology and human-computer interaction technology, which can provide theoretical reference for subsequent related research.

Keywords: HMM; spoken English; training system; feature recognition; oral training

DOI: <https://doi.org/10.14733/cadaps.2020.S1.101-112>

1 INTRODUCTION

Spoken language is an important tool for achieving communication and communication between people, and the proficiency in spoken English is an important criterion for measuring conversational ability. With the implementation of the new curriculum reform, China's current English teaching has increasingly paid attention to the cultivation of communicative competence such as English listening and speaking, and more and more English learners are paying more and more attention to their ability in spoken English and spoken English teaching. Moreover, spoken English teaching has become one of the important research fields of applied language discipline [1]. In today's information age, the rapid development of computer technology, network technology and speech recognition technology has solved the biggest problem in oral training application from the environment. Moreover, computer-aided language learning system (CALL) has become an optimal oral learning

method [2]. This learning method, which breaks through the limitations of time and space, allows students to train anytime and anywhere, and can relieve the psychological pressure of students to be afraid of making mistakes. After the students are trained, the system can also feedback the results of the exercises in a timely and accurate manner, so that students can quickly find their weaknesses and can take complementary learning in a targeted manner. Through a large number of exercises in the simulated real environment, the level of students' speaking ability will be greatly improved [3].

Speech recognition technology, also known as Automatic Speech Recognition (ASR), is one of the ten most important technological development technologies in the information technology field from 2000 to 2010. Speech recognition technology is the key technology of human-computer interaction in information technology, and it is mainly used for input of human-computer interaction. The goal is to allow the machine to receive voice messages from people and convert them into computer-readable inputs, such as buttons, binary codes, or sequences of characters. It is equivalent to installing an "ear" on a computer system, so that the computer has the function of "listening". Moreover, it allows the computer to convert the voice signal into corresponding text or command through the process of recognition and understanding, thereby realizing the most natural and convenient means of "voice" in the information age for human-machine communication and interaction [4].

AT&T Bell Labs Davis et al. successfully developed the world's first experimental system that effectively recognizes the pronunciation of 10 English numerals, the Audry system. The identification method of the system is to track the formant in the speech, and the correct rate of 98% can be obtained [5].

Artificial neural networks were introduced into speech recognition, and two advanced technologies, dynamic time warping (DTW) and linear predictive coding (LPC), were proposed for speech signals. The proposed two techniques provide an effective solution to the problem of phonological feature extraction and speech signal matching unequal length [6]. The rapid development of these technologies has provided a good environment and laid a good foundation for the development of speech recognition technology and even the entire intelligent speech technology. Speech recognition technology had the most significant breakthrough, that is, the hidden Markov model (HMM) was applied. In the basic of dynamic time warping (DTW) and linear predictive coding (LPC), based on the principle of pattern matching, for a specific person, an isolated word based on a small vocabulary is recognized, thereby realizing a speech recognition system for a specific person's isolated words. At the same time, the related theories of speech signal vector quantization (VQ) and hidden Markov model (HMM) are also proposed. The study of speech recognition has made substantial progress in the identification of isolated words and small vocabulary sentences [7].

Strictly speaking, speech recognition technology did not break away from the HMM framework, but with the expansion of the application field, laboratory speech recognition research broke through the three major obstacles of large vocabulary, continuous speech and non-specific people. The first system to integrate these three features into one system was the Sphinx system at Carnegie Mellon University. It is the first high-performance recognition system that meets non-specific, large vocabulary and continuous speech [8].

Speech recognition research has been further advanced. Its distinctive feature is the successful application of HMM model and artificial neural network (ANN) in speech recognition. The application of artificial neural network technology in speech recognition has opened up a new way for speech recognition. It enables speech recognition to have parallel, linear characteristics, and it also makes it more adaptive, fault tolerant, and self-learning [9]. It can be said that the introduction of artificial neural network technology provides a great impetus for the development of speech recognition technology, and also enables the speech recognition system to be truly used from the laboratory. At present, speech recognition technology has made great progress in productization and application. Many developed countries such as the United States, Japan, South Korea, and famous companies such as IBM, Apple, AT&T, and NTT have adopted the practical development of speech recognition systems [10].

The more representative systems that the research has invested heavily are: IBM launched ViaVoice and Dragon System's Naturally Speakin, Nuance's Nuance Voice Platform, Microsoft's Whisper, Sun's VoiceTone, etc. [11]. Moreover, in the developed countries of the western economy, a large number of speech recognition products have entered the market and service fields. For example, some users' telephones and mobile phones already have voice recognition dialing functions, and some voice memos and smart voice toys are also embedded with voice recognition and voice synthesis. In addition, people can also find air tickets, banks, and travel information in the dialogue system through voice recognition in the telephone network. According to statistics, more than 85% of users are satisfied with their performance [12].

2 RESEARCH METHODS

2.1 Traditional scoring algorithm

The requirements for the scoring algorithm are: (1) Better reliability and consistency with expert ratings. (2) It reflects the learner's ability to pronounce and does not pursue the best similarity between it and the standard pronunciation individual.

After research, it is found that the HMM-based phoneme posterior probability algorithm has very good stability and is not easily affected by the individual characteristics of the learner or the changes of the sound channel and has better feedback on the similarity between the learner's pronunciation and the standard pronunciation [13].

During speech processing, input features and templates cannot be directly compared directly because the speech signal is very random. Even if the same person reads the same sentence, it is impossible to have the exact same length of time. For example, as the vocalization speed increases, the length of time of the vowel stable portion will be shortened, while the length of the consonant or transitional sound portion remains substantially unchanged. Therefore, time regulation is essential. Dynamic time warping is a nonlinear regularization technique that combines regular time and measure distance calculation. Hypothesis: The feature vector sequence of the reference template is $a_1, a_2, \dots, a_m, \dots, a_M$, and the sequence of feature vectors of the input speech is $b_1, b_2, \dots, b_n, \dots, b_N$, and $b_1, b_2, \dots, b_n, \dots, b_N$. Thus, dynamic normalization is to find a time warping function $m = w(n)$ that nonlinearly maps the time axis n to the reference template time axis m , such that [14].

$$D(n, w) = \min_{w(n)} \sum_{n=1}^N d[n, w(n)] \quad (1)$$

In the above formula, $d[n, w(n)]$ represents the distance between the feature vector of the n th input and the $w(n)$ reference template vectors. Obviously, $w(n)$ should be a non-decreasing function. In time, dynamic time alignment aligns the input features and reference template features to eliminate unnecessary differences between the two. Figure 1 shows a schematic diagram of the magnitude of the distortion between linear matching and nonlinear matching, direct matching. The nonlinear matching method shown in the figure is likely to minimize the non-essential difference between the two modes [15].

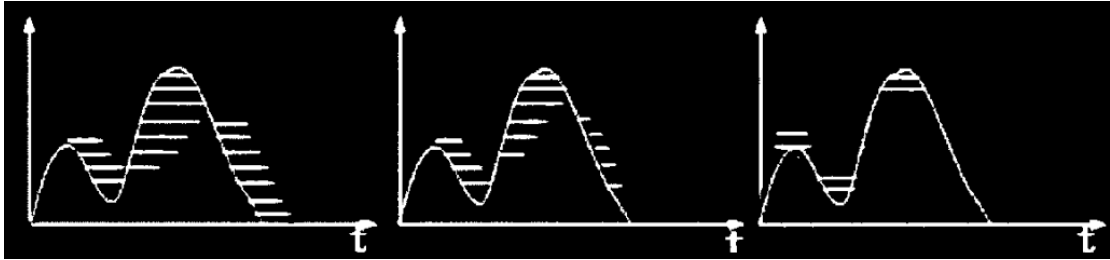


Figure 1: Pattern matching diagram.

Dynamic time warping is an optimization problem. A common dynamic programming technique for solving this problem takes advantage of the concept that local best results in the overall optimal value. The goal of the solution is to find the optimal time warping function $w(n)$ and the corresponding $D(n, w)$. In some specific problems, the DTW function satisfies the following conditions.

Boundary conditions [16]:

$$w(1) = 1, w(N) = M \quad (2)$$

Continuous conditions:

$$w(n+1) - w(n) = \begin{cases} 0, 1, 2 & w(n) \neq w(n-1) \\ 1, 2 & w(n) = w(n-1) \end{cases} \quad (3)$$

A recursive formula can be introduced:

$$D(n+1, m) = d[n+1, m] + \min[D(n, m)g(n, m), D(n, m-1), D(n, m-2)] \quad (4)$$

Among them:

$$g(n, m) = \begin{cases} 1 & w(n) \neq w(n-1) \\ \infty & w(n) = w(n-1) \end{cases} \quad (5)$$

Since the calculation of each $D(n+1, m)$ requires calculation of the D values of the three points on the n columns, the use of dynamic programming techniques is very time consuming when calculating time warping. Pattern recognition often calculates the distance between features. In speech recognition, the degree of similarity between the reference mode and the input mode is determined by the degree of distortion between the frames that make up the two. It is a measure that reflects the difference in signal characteristics and is expressed as $D(x, y)$. This is the applied distance measurement method that needs to satisfy the following mathematical properties [17]:

Positive value: $D(x, y) \geq 0$. When $x = y$, there is $D(x, y) = 0$.

Symmetry: $D(x, y) = D(y, x)$

Triangle inequality: $D(x, y) \leq D(x, z) + D(z, y)$

When calculating the DTW distance, the absolute average distance is used:

$$D(x, y) = \frac{\sum_{i=1}^N |x_i - y_i|}{N} \quad (6)$$

DTW distances are not directly categorized as pronunciations, and a reasonable mapping from distance to score must be sought. Assuming that the relationship between distance and score satisfies the formula [18]:

$$score = \frac{100}{1 + a(dist)^b} \quad (7)$$

Obviously, this formula can map distances to a range of 0 to 100. According to this, the distance-score conversion is shown in Figure 2.

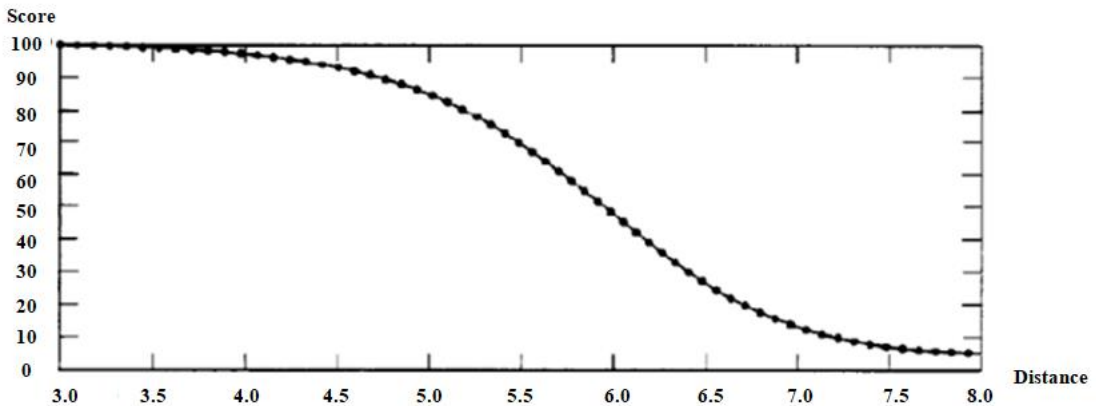


Figure 2: Distance-score conversion chart.

When solving the unknown parameters a , b in the formula, we need to know some score and distance pairs. The above parameters can be solved by the scores of some experts in the experiment and the DTW distance. When using the formula of this paper, even if the distance is larger or smaller than the test, the score can be reasonably converted to the range of 100 to 0. Since the two feature parameters are actually used, the actual score estimation formula is slightly complicated, and the final score is the weighted sum of the two.

$$score = w_1 * \frac{100}{1 + a_1(dist_1)^{b_1}} + w_2 * \frac{100}{1 + a_2(dist_2)^{b_2}} \quad (8)$$

The parameters in the formula satisfy these constraints: $a_1, a_2, b_1, b_2 > 0, w_1 + w_2 = 1$. a_1, a_2, b_1, b_2 is the parameter of the distance component, and w_1, w_2 is the weight of the three features.

2.2 HMM-based scoring method

As shown in Figure 3, it is the system's scoring system process. The standard answer is based on the pre-trained acoustic model and tone model. Among them, speech recognition technology is used to find the difference between the test speech and the model and combined with the scoring mechanism to give a score. The feature parameter extraction mainly includes two characteristic parameters: the fundamental frequency trajectory and the Mel cepstral parameter, which are used as the characteristic parameters of tone recognition and sound recognition respectively. The use of Viterbi Decoding first divides the speech signal into a single syllable. Then, the sound model and the

tone model are compared for each syllable, and the recognition results are combined with our pre-designed scoring mechanism to convert the scores, that is, the scores of the test speeches. This scoring system includes the Hidden Markov Model (HMM), the TreeNet and the Viterbi Algorithm, which are commonly used in speech recognition. In terms of tone recognition, Chebyshev Approximation, Orthogonal Expansion, K. means grouping method and class design are included. There is no spanning HMM structure from left to right as shown in Figure 4.

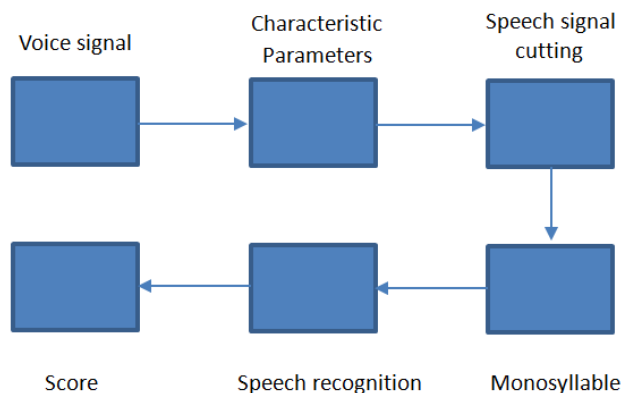


Figure 3: Schematic diagram of the scoring system.

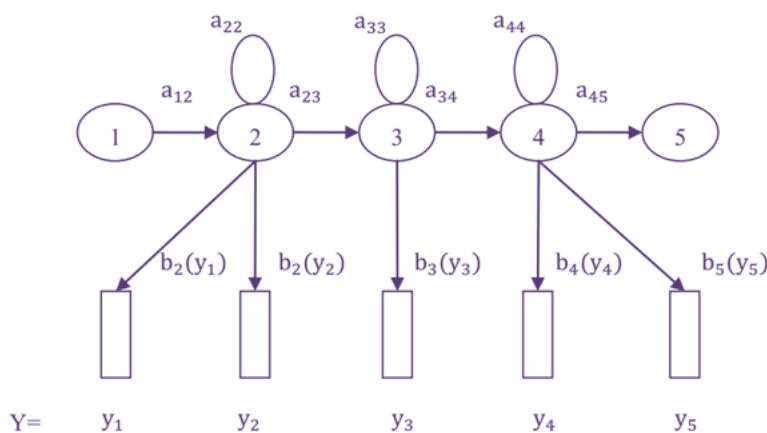


Figure 4: No cross-over HMM structure diagram from left to right.

The difference between the HMM method and the DTW method is: The HMM pattern library is not a pre-stored pattern sample, but a set of H/VIM model parameters with the highest probability of combining with the training output signal formed by repeated training and iterative algorithm (Baum-Welch). A is the state transition probability distribution matrix, and B is the system output probability distribution matrix. The parameters reflect the numerical parameters of the statistical properties of the training speech process, not the mode feature parameters themselves. HMM-based scoring is a statistical model-based approach that differs from DTW algorithm template matching.

The HMM-based method is mainly used for the scoring of phonemes. The more common methods are scoring of log likelihood and scoring of logarithmic posterior probability. Compared to feature comparison scores, this method reflects the learner's ability to pronounce a language, not just the difference between a standard speaker individual.

For the log likelihood score, the formula is defined as follows:

$$S_i = \sum_{t=\tau_j}^{\tau_{j+1}} \lg [P(q_t | q_{t-1}) P(o_t | q_t)] \quad (9)$$

Among them, o_t and q_t are the observation vector and the state of HMM at time t , respectively. If the definition of the model is $\lambda = (M, N, A, B, \pi)$, $P(q_t | q_{t-1})$ is the state transition probability, that is, A in the HMM model. Among them, $P(o_t | q_t)$ is the probability distribution matrix of the observation vector, that is, B in the HMM.

The scoring method for scoring sentences is:

$$S = \sum W_i S_i / \sum W_i \quad (10)$$

Among them, S is the sentence score, Log likelihood scores can be used for both text-related and text-independent situations. After analyzing the characteristics of several pronunciation algorithms, the system uses the HMM-based phoneme posterior probability algorithm as a reference for the pronunciation evaluation algorithm.

The score based on the HMM posterior probability is:

Among them, $P(O_t | q)$ is the probability distribution of the observation vector O_t under the phoneme q , and $P(q)$ is the prior probability of the phoneme q . The sum on the denominator is the sum of the phonemes $q=1, \dots, M$ that are independent of all the text.

The phoneme q_i takes the logarithm of the posterior probability of each pause in the i -th segment of speech, and then the obtained value is accumulated to obtain the logarithmic posterior probability score of the phoneme q_i under the i -th segment of speech:

$$P_i = \sum_{t=\tau_j}^{\tau_{j+1}} \lg [P(q_t | o_t)] \quad (11)$$

Because the pronunciation of English beginners is slow, the speech rate is also a factor that affects the pronunciation score. The formula for the definition of the phoneme duration is:

$$D = \frac{1}{N} \sum_{i=1}^N \lg [P(f(d_i | q_i))] \quad (12)$$

d_i is the duration of the i -th segment of phoneme q_i and is a normalized function. Taking into account the independence of text and scholars, the speech rate (ROS) metric is used to normalize the speech duration. ROS is the number of phonemes per unit of time in a sentence or speaker's utterance, which is usually taken as $f(d_i) = ROS * d_i$.

Under normal circumstances, the primary English practitioner's speech postponement probability score will be lower than that of the English professional speaking teacher for training. This rating reflects to some extent certain errors in the pronunciation of non-native English speakers.

2.3 Error detection

After the MFCC feature value is subjected to the forced correlation recognition and scoring process of the phoneme, the corresponding associated phoneme string, phoneme start time and end time score are obtained. After these results were obtained, we started the phoneme error detection.

According to the results of the most similar phoneme judgment, the phonetic reading errors are roughly divided into three categories: misreading, missing reading, and adding phonemes. The most similar phoneme is defined as the phoneme with the highest HMM likelihood:

$$q_i = \arg \text{Max}[L_i(q)] \quad (13)$$

$L_i(q)$ is the likelihood of any one of the factors g in the i period:

$$L_i(q) = P(q|Q_i) = P_i = \sum_{t=\sigma_i}^{\sigma_{i+1}-1} 1g [P(s_t|s_{t-1})P(o_t|s_t)] \quad (14)$$

Missed phoneme: The tone of q_i is not sent.

$$q_i = \begin{cases} q_i \\ q_{i+1} \\ SIL \end{cases} \quad (15)$$

Misreading phonemes: The pronunciation of q_i is incorrect and sounds more like other pronunciations. It is expressed as $q_i \neq q_i$ and is not a missed phoneme error.

Add phoneme: The recognition result contains extra phonemes. The results obtained by the correlator cannot distinguish between these three types of errors. Among them, the error detection module can only locate the wrong phoneme according to the phoneme score. For further research, it is necessary to obtain the identified phoneme result by the recognition process to determine which kind of error is specific. For different needs, we can design two methods of error detection.

If the type of error detection is not required, we first use the correlator to score the evaluation voice to correlate the phoneme level and set the threshold. When the corresponding phoneme score is below the threshold, the phoneme is included in the wrong phoneme. If the specific type of error needs to be detected, an additional phoneme recognition process is required.

3 SYSTEM DESIGN

The modules of the speech recognition system mainly include five parts: feature value extraction, phoneme recognition, phoneme association, pronunciation evaluation and error detection. Finally, the system will give the learners effective feedback results in terms of pronunciation scores and corrective opinions. This learning method enables learners to better understand the pronunciation errors and correct them to improve the oral level. The speech recognition module is shown in Figure 5.

External function: Implement a visualization window. Example library view: Follow the specified example or learning strategy during playback. Information view: Display sentences and phonetic symbols; display single phoneme scores and corrections; display sentence prosody scores and corrections, play corrections. User Management Interface: Display user's vocabulary, learn files, analyze misreading phonemes and common mistakes. The system recognizes English learners with a strong Chinese accent. The system designed in this paper is implemented by C++ programming and runs on the Windows platform. A small part of the system is written in MATLAB. Specific requirements are as follows: (1) The occupied memory space is small. (2) the limitation of the operating system to the operating system is limited. (3) The realization of English-based EI training system based on HMM under VC platform needs to properly consider the scalability of the system itself and the friendliness of the interface. (4) For a given user pronunciation, the system can give the user a matching score, and finally provide feedback to the user to help learn English progress.

For the voice processing diagram depicted in Figure 4, the system is divided into three parts: user interface GUI, input/output I/O, and Scorer. The system module relationship is shown in Figure 6.

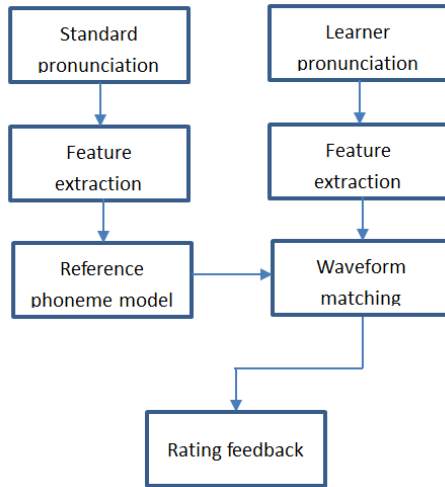


Figure 5: Schematic diagram of the speech recognition module.

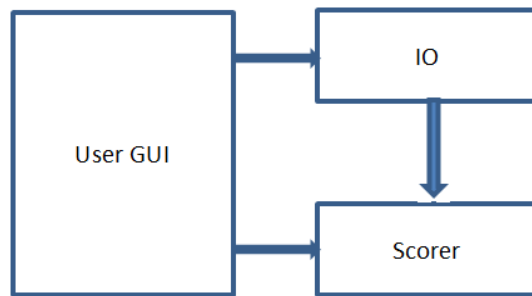


Figure 6: Relation chart of system module diagram.

The main interface design effect is shown below:



Figure 7: Renderings of interface.

The speech segmentation effect diagram after the boundary modification is as follows:

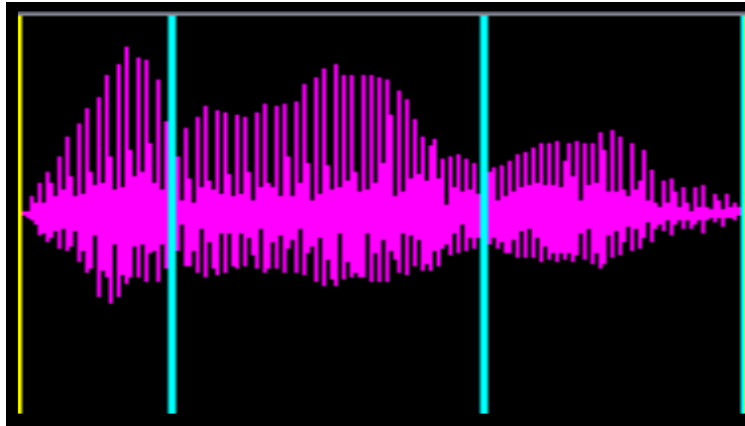


Figure 8: Speech segmentation effect map after boundary modification.

4 ANALYSIS AND DISCUSSION

The evaluation of speech quality is not only related to the disciplines of linguistics, phonetics, signal processing, but also related to science, psychology, and even cultural traditions. It is a rather complicated aspect. The method of speech quality evaluation is based on subjective scoring and objective scoring. There are many common scoring methods, such as: distortion average opinion score, average opinion score, diagnostic rhyme test method, judgment satisfaction test method, and so on. For the subjective scoring of speech, its shortcomings are time-consuming and laborious, and because of the limitations of many test conditions and the subjective factors of the testers, the reliability of the test results is affected to some extent. Therefore, the objective evaluation of the quality of speech usually requires some equipment, which is flexible and flexible, and is not affected by some realistic conditions and artificial factors. At the same time, we can directly compare test results at different times and different occasions. So far, many methods for evaluating the quality of pronunciation have been developed. Our commonly used methods are: score based on the HMM log similarity, score based on the dynamic time regular DTW, and score based on the HMM logarithm posterior probability, segment time score, segment classification score, fluency score and other methods. The above scoring methods are obtained by performing standard matching of various effective similarities according to the standard pronunciation as a reference template.

DTW is an optimization algorithm. In order to align its features and template features, the time axis of the speech signal to be recognized is unevenly warped and curved, and the matching path between the two vector distances is calculated directly between the two. Therefore, the earliest and most commonly used method to successfully solve the speech pattern matching problem is to obtain the regularization function with the smallest accumulation distance when the two vectors match. The disadvantages of the DTW method are that the amount of computation is large, the number of endpoint detections of the speech signal is too large, and the timing dynamic information of the speech signal is not sufficiently utilized.

The requirements for the scoring algorithm are: (1) Better reliability and consistency with expert ratings. (2) It reflects the learner's ability to pronounce and does not pursue the best similarity between it and the standard pronunciation individual.

After research, it is found that the HMM-based phoneme posterior probability algorithm has very good stability and is not easily affected by the individual characteristics of the learner or the changes of the sound channel and has better feedback on the similarity between the learner's pronunciation and the standard pronunciation. The modules of the speech recognition system mainly include five parts: feature value extraction, phoneme recognition, phoneme association, pronunciation evaluation

and error detection. Finally, the system will give the learners effective feedback results in terms of pronunciation scores and corrective opinions. This learning method enables learners to better understand the pronunciation errors and correct them to improve the oral level.

5 CONCLUSION

Aiming at the difficulty of oral English teaching in current English teaching, this paper used computer technology, network technology, natural language processing technology, speech processing technology and human-computer interaction technology to exploratorily design a spoken language training system based on automatic speech evaluation. The system solves the biggest problem in oral training application from the environment, provides a learning method that can break through time and space constraints, and allows students to train anytime and anywhere, and can alleviate the psychological pressure of students afraid of making mistakes. The research field of computer-aided language learning system based on speech technology is still in its infancy, and speech recognition is also a hot research direction of speech technology. Moreover, research on automatic speech evaluation based on speech recognition has been carried out in full swing. At present, the application of automatic voice evaluation in the oral training system has enabled the learner to accurately understand his own speaking ability and can locate the pronunciation error and give the user a corrective advice. However, in the current evaluation results, the positioning and correction of the discourse is not very accurate and perfect. Therefore, how to detect and locate the pronunciation errors more accurately and give correction suggestions is the direction of further research by the project team in the future.

6 ORCID

Hui Li, <https://orcid.org/0000-0002-7514-6712>
 Xinyu Zhang, <https://orcid.org/0000-0003-1775-5218>
 Ran Cui, <https://orcid.org/0000-0001-7719-8148>
 Na Wang, <https://orcid.org/0000-0001-5767-0732>

REFERENCES

- [1] Dat, N. D.; et al: STING algorithm used English sentiment classification in a parallel environment, *International Journal of Pattern Recognition and Artificial Intelligence*, 31(07), 2017, 25, <https://doi.org/10.1142/S0218001417500215>
- [2] Darabseh, A.; Siami-Namini, S.; Siami Namin, A.: Continuous authentications using frequent english terms, *Applied Artificial Intelligence*, 2018, 1-35, <https://doi.org/10.1080/08839514.2018.1447535>
- [3] Xia, C.; Jing, Z.: English translation of classical Chinese poetry, *Orbis Litterarum*, 73(4), 2018, <https://doi.org/10.1111/oli.12184>
- [4] Lu, Z.; et al: Design of voice interaction-based training system for cerebral palsy rehabilitation, *Chinese High Technology Letters*, 2017, <https://doi.org/10.3772/j.issn.1002-0470.2017.03.011>
- [5] Köster, F.; et al: Towards degradation decomposition for voice communication system assessment, *Quality & User Experience*, 2(1), 2017, 4, <https://doi.org/10.1007/s41233-017-0006-5>
- [6] Singh, S.; et al: Information communication technology for extension: a mobile phone based voice call system for dissemination of cotton production technologies, *Journal of Agricultural & Food Information*, 2018, 1-9, <https://doi.org/10.1080/10496505.2018.1436442>
- [7] Gundogdu, K.; Bayrakdar, S.; Yucedag, I.: Developing and modeling of voice control system for prosthetic robot arm in medical systems, *Journal of King Saud University - Computer and Information Sciences*, 2017, S1319157817300216, <https://doi.org/10.1016/j.jksuci.2017.04.005>

- [8] Calder, L. A.; et al: The feasibility of an interactive voice response system (IVRS) for monitoring patient safety after discharge from the ED, *Emergency Medicine Journal*, 2017:emermed-2016-206192, <https://doi.org/10.1136/emermed-2016-206192>
- [9] Carrie, E.; Mckenzie, R.: American or British? L2 speakers' recognition and evaluations of accent features in English, *Journal of Multilingual & Multicultural Development*, 2018, 313-328, <https://doi.org/10.1080/01434632.2017.1389946>
- [10] Boutsen, F. R.; Dvorak, J. D.; Deweber, D. D.: Prosody and spoken word recognition in early and late Spanish-English bilingual individuals, *Journal of Speech Language and Hearing Research*, 60(3), 2017, 712, https://doi.org/10.1044/2016_JSLHR-H-15-0274
- [11] Sidgi, L. F. S.; Shaari, A. J.: The effect of automatic speech recognition Eyespeak software on Iraqi students' English pronunciation: A pilot study, *Advances in Language and Literary Studies*, 8(2), 2017, 48, <https://doi.org/10.7575/aiac.all.v.8n.2p.48>
- [12] Zou, S.: Designing and practice of a college English teaching platform based on artificial intelligence, *Journal of Computational and Theoretical Nanoscience*, 2017, <https://doi.org/info:doi/10.1166/jctn.2017.6133>
- [13] Fragkou, P.: Applying named entity recognition and co-reference resolution for segmenting English texts, *Progress in Artificial Intelligence*, 6(3), 2017, 1-22, <https://doi.org/10.1007/s13748-017-0127-3>
- [14] Zhang, B.; Xiong, D.; Su, J.: Neural machine translation with deep attention, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 1-1, <https://doi.org/10.1109/TPAMI.2018.2876404>
- [15] Zhang, H.; et al: Understanding Subtitles by Character-Level Sequence-to-Sequence Learning, *IEEE Transactions on Industrial Informatics*, 13(2), 2017, 616-624, <https://doi.org/10.1109/TII.2016.2601521>
- [16] Stahlberg, F.; Byrne, B.: Unfolding and shrinking neural machine translation ensembles, 2017, <https://doi.org/10.18653/v1/D17-1208>
- [17] KlubiKa, F.; Toral, A.; Sánchez-Cartagena, V. M.: Quantitative fine-grained human evaluation of machine translation systems: a case study on English to Croatian, *Machine Translation*, 2018, <https://doi.org/10.1007/s10590-018-9214-x>
- [18] Arcan, M.; et al: Leveraging bilingual terminology to improve machine translation in a CAT environment, *Natural Language Engineering*, 23(5), 2017, 26, <https://doi.org/10.1017/S1351324917000195>