Frontiers in Educational Research
ISSN 2522-6398 Vol. 1, Issue 2: 10-16, DOI: 10.25236/FER.030008

Application of Speech Recognition Technology on the Evaluation of English Pronunciation Teaching

Fanyu Wang

Baotou Vocational &Technical College,Inner Mongolia, Baotou,014035,China

ABSTRACT. *In order to better the current English learning environment and teaching mode so as to improve the efficiency of spoken English learning, the speech recognition technology is applied to the scoring of English pronunciation teaching. It is pointed out that the feature score be divided into three domains, including the pronunciation segment, the hyper articulation segment and the perceptual domain. The speech recognition technology is used to identify the pronunciation features, and the value of recognition accuracy is used to get the fractional value. After calculating and measuring the correlation coefficient, the correlation coefficient between the synthesized machine score by the three domains and expert score is higher than that of the pronunciation segment score, which is also higher than the effect of the synthesized machine score of any two fields. The research shows that the performance of the scoring mechanism is much better than the previous scoring mechanism, suggesting that it is helpful to the evaluation of English pronunciation teaching.*

KEYWORDS: *Speech recognition technology; English pronunciation teaching; pronunciation section; scoring mechanism.*

1. Introduction

The rapid development of computer technology brings the opportunity to the reform of English teaching. Language leaming using computer (CALL, Computer Aided Language Leaming) has become the inevitable trend of College Teaching Reform in information age, as a way of the best oral English learning. The evaluation of pronunciation quality, that is, speech scoring mechanism is the key technology in CALL system. It is a computer that uses speech processing technology to correctly evaluate learners' pronunciation exercises instead of evaluate by experts [1]. At present, most of the researches on scoring mechanism of CALL system are all about extracting the acoustic characteristics of speech signals, namely, the evaluation of speech segment features, which is obviously not comprehensive enough, ignoring the information hidden in other aspects of speech signals. Therefore, a better scoring method and scoring mechanism should be explored to achieve

ISSN 2522-6398 Vol. 1, Issue 2: 10-16, DOI: 10.25236/FER.030008

a more accurate evaluation of learners' speech, which helps learners find problems in their pronunciation and better promote the efficiency of spoken English learning.

Speech scoring technology is used to determine the accuracy of the pronunciation of the speaker. If the scoring is to be done by computer, the support of speech processing technology will be needed, and the key is speech recognition technology [2]. In recent years, with the improvement of computer speed and the progress of speech recognition technology, the evaluation of spoken speech by using automatic speech recognition technology has become a research hotspot. Study abroad in the oral assessment has been very in-depth and involves in many aspects of oral features: evaluated from the statement, word or phoneme and other different levels of spoken English pronunciation. Some systems are also able to produce the similar results with language expert's evaluation results. At home, this research has just started. This paper studies the application of speech recognition technology in English pronunciation teaching. It also provides reference and help in this field.

2. Methodology

2.1 Speech recognition technology

Figure 1 is a block diagram of the principle of an automatic speech recognition system.

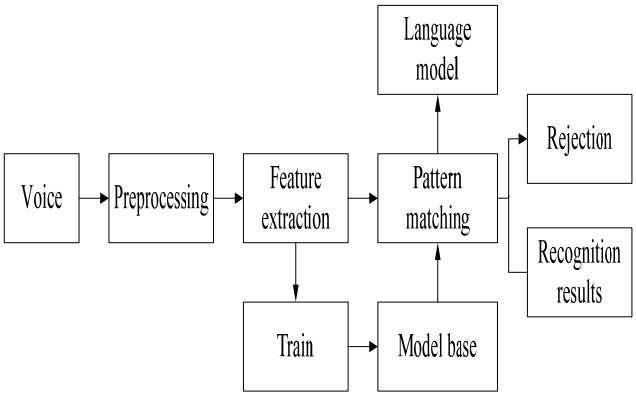

Fig.1 Block diagram of the principle of an automatic speech recognition system

Feature template training is a key part for speech recognition, when identify a speech, speech template library needs to have a standard, to match the identified voice and the standard template library. The maximum likelihood standard speech is set as matching results. This template library needs to be trained by a large number of samples [3]. In the experiments the HMM model based on template training is used, because the implicit Maerkekefu model is a statistical model based on the technology which can well describe the variability and stability of speech signals. It can integrate the advantages of acoustics and statistical knowledge from linguistics to syntactic into

ISSN 2522-6398 Vol. 1, Issue 2: 10-16, DOI: 10.25236/FER.030008

a unified framework.

2.2 Scoring Technology

As a statistical model of speech signal, hidden Markov model (HMM) is widely used in all fields of speech processing. Hidden Markov model is a kind of Markov chain, and the reality is more complex. The observed events do not correspond to the state one by one but are linked by a group of probability distributions. Such a model is called HMM. It is a double random process, one of which is the Markov chain, which is the basic random process, which describes the transfer of states. Another random process describes the statistical correspondence between the state and the observed value. What observers can see is only the observation value, not the state, but through a random process the existence and characteristics of the state can be perceived. The study of speech uses the principle of HMM [4].

Considering the process of speech recognition, in the period of user training or recognition, even if every time describe one word or one sentence as far as possible in the same way, the length of duration will be changed randomly, and each word inside the various parts (such as vowels and consonants) relative to the market is a random change. The need to align with the characteristic parameters of time sequence pattern again calls for the adoption of the dynamic time warping (DTW) method, which can solve this problem effectively. It is also a very successful matching technique in speech recognition, [5].

The machine score is evaluated by the computer's pronunciation of learners, and the score obtained is relative to people's subjective score, which is objective [6]. At present, many scoring techniques have been widely used, including the HMM likelihood log score and the HMM log posterior probability score.

The HMM log likelihood score is the likelihood logarithm of the spectral observation value extracted from the short time window of the speech as a score. The potential assumption of this method is that the likelihood log of speech data calculated by Viterbi algorithm is a good way to measure native and non-native speech, and the premise is that HMM model is trained with native voice data.

$$l_i = \sum_{t=\tau_i}^{\tau_{i+1}-1} \log\left(p\left(s_t \mid s_{t-1}\right) p\left(x_t \mid s_t\right)\right) \tag{1}$$

HMM's log posterior probability score is robust, and it is not easy to be changed by learners' individual characteristics or vocal tract changes, which can better reflect the similarity between learners' pronunciation and quasi pronunciation.

$$p\left(q_t \mid x_t\right) = \frac{p\left(x_t \mid q_t\right) p\left(q_t\right)}{\sum_{q=1}^{M} p\left(x_t \mid q\right) p\left(q\right)} \tag{2}$$

ISSN 2522-6398 Vol. 1, Issue 2: 10-16, DOI: 10.25236/FER.030008

$$p_i = \sum_{t=\tau_i}^{\tau_{i+1}-1} \log\left(p\left(q_t \mid x_t\right)\right) \tag{3}$$

$$P = \frac{1}{N}\sum_{i=1}^{N}\frac{P_i}{d_i} \tag{4}$$

An expert scoring is a subjective process, in which a true and reliable evaluation of the voice of the test is given. In the research of Bernstein and Greenberg's VILTS project, the experts' evaluation of student's voice is divided into 7 grades. Grade 7 indicates that the pronunciation level of students is very good, basically reaching the standard level. Grade L indicates that students' pronunciation has heavy accent, even difficult to understand. The expert rating system used in this paper is slightly different from the expert grading system developed in VILTS research. This paper decided to adopt a 1-5 level instead of l-7 level. Because the level of the latter is more than the former, it is more likely to lead to the inconsistency of the result, the low correlation between the experts, making it not so convincing. Therefore, the 1-5 level system is used and corresponds each to the corresponding scores, as shown in the following table:

Table 1: Scoring levels set by experts

Levels	Pronunciation performance	Scores
1	Very good	5
2	Good	4
3	Not too bad	3
4	Bad	2
5	The worst	1

An expert score is an average of five experts on the overall score of a sentence, for example:

$$S_e = \frac{1}{n}\sum_{i=1}^{n} S_i \tag{5}$$

2.3 THE DESIGN OF THE SCORING MECHANISM

In this paper, two speech databases are needed, one is the standard training speech database, which is the standard acoustic model used for training, the other is the speech library to be tested, and it is used for testing and grading. The standard training speech database adopts the International English continuous speech database TIMIT, which is composed of 630 voice data from 8 main dialects in the United States. A total

ISSN 2522-6398 Vol. 1, Issue 2: 10-16, DOI: 10.25236/FER.030008

of 100 sounds, recorded by a man and a woman, 50 sentences per person. The text of the voice file is also a copy from the TIMIT corpus [7].

The acoustic model should be established by the HMM model of mono-phoneme. The HMM models of 62 phonemes all use the same topology. Then the list of three phonon models is generated by the HMM model of the mono phoneme. The structure of the three tone is a phoneme-phoneme + phoneme. The semantic network is implemented by the HParse tool in the HTK tool [8]. In this paper, the score of learners' voice is calculated from three aspects, namely, articulation, hyper voice and perception domain. Therefore, we need to extract different representative speech features for these three aspects. The window processing is used here, using Hanming window.

$$\omega(n)=\begin{cases}0.54-0.46\cos\left[2\pi n/(N-1)\right], 0\le n\le N-1\\ 0, others\end{cases} \tag{6}$$

In this experiment the pronunciation, speech from Mel cepstral coefficients (MFCC) serves as a representative feature, first of all, using HTK tools to extract the 4620 sound files in the standard library and the MFCC features 100 sound files in the voice database to be measured, and using 4620 MFCC feature vector the standard voice to the established three tone sub training model, generating a standard acoustic model library. This score is the score of suprasegmental features from the view of phonology, is mainly from the perspective of rhythm, including stress and intonation. In this paper, the loudness of speech signals and the two features of RASTA-PLP are extracted from the perceptual domain information of speech signals to be used to score the perceptual domain characteristics of the speech signals.

3. Results and discussion

Whether the machine score obtained by computer can really replace the expert scoring, it is needed to evaluate the performance of machine scoring. This paper uses machine score and expert score correlation coefficient to describe:

$$r=\frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum\left(x_i-\bar{x}\right)^2\sum\left(y_i-\bar{y}\right)^2}} \tag{7}$$

The correlation coefficient between the score from the machine for the fusion of three domains and the expert is shown in Table 2 as follows:

TABLE 2: CORRELATION COEFFICIENT OF LINEAR SYNTHETIC MACHINE SCORE AND EXPERT SCORING

Combining approach	Related efficient
Scores of segmental + suprasegmental	0.686

ISSN 2522-6398 Vol. 1, Issue 2: 10-16, DOI: 10.25236/FER.030008

Scores of segmental +perception domain score	0.693
Suprasegmental +perception domain score	0.622
Scores of segmental + suprasegmental +perception domain score	0.733

Each characteristic score of the speech to be measured is used as the input set of the network, and the expert score of the speech to be measured is used as the output set of the network, and the BP neural network is trained. Table 3 is the comparison of the correlation coefficients between the score of the machine and the expert score by the different synthetic methods of the BP neural network:

TABLE 3: CORRELATION COEFFICIENT RESULTS OF BP NEURAL NETWORK SYNTHETIC MACHINE SCORE AND EXPERT SCORING

Combining approach	Related efficient
Scores of segmental + suprasegmental	0.753
Scores of segmental +perception domain score	0.744
Suprasegmental +perception domain score	0.652
Scores of segmental + suprasegmental +perception domain score	0.815

From table 2 and table 3, it can be seen that the correlation between machine score and expert score obtained by combining three domain feature scores is higher than that of other synthetic methods. Moreover, using training BP neural network to synthesize the three-domain feature score, the correlation coefficient between the machine score and the expert score can reach 0.815 when the best score is achieved, which is higher than the linear synthesis method and achieves the expected effect.

4. Conclusion

According to the difficult problems in English speech teaching, the computer assisted language learning system scoring mechanism is expounded in the paper. Exploratory design of a scoring mechanism based on speech recognition technology is made. It is suggested to evaluate from the characteristics of the three domains of a speech, and then use artificial neural network to get a synthesis a total score of the machine. The experimental results show that the correlation coefficient between the score from the machine and the score given by expert has been greatly improved.

References

[1] Li, J., Deng, L., Gong, Y. (2014). An Overview of Noise-Robust Automatic Speech Recognition, IEEE/ACM Transactions on Audio Speech & Language Processing, Vol.22, No.4, pp.745-777.

[2] Waibel, A. (2014). Modular Construction of Time-Delay Neural Networks for Speech Recognition, Neural Computation, Vol.1, No.1, pp.39-46.
[3] Besacier, L., Barnard, E., Karpov, A. (2014). Automatic speech recognition for under-resourced languages: A survey, Speech Communication, Vol.56, No.1, pp.85-100.
[4] Saon, G., Kuo, H. K. J., Rennie, S. (2015). The IBM 2015 English Conversational Telephone Speech Recognition System, Eurasip Journal on Advances in Signal Processing, Vol.2008, No.1, pp.1-15.
[5] Kim, C., Stern, R. M. (2016). Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition, IEEE/ACM Transactions on Audio Speech & Language Processing, Vol.24, No.7, pp.1315-1329.
[6] Swietojanski, P., Ghoshal, A., Renals, S. (2014). Convolutional Neural Networks for Distant Speech Recognition, IEEE Signal Processing Letters, Vol.21, No.9, pp.1120-1124.
[7] Healy, E. W., Yoho, S. E., Wang, Y. (2013). An algorithm to improve speech recognition in noise for hearing-impaired listeners, Journal of the Acoustical Society of America, Vol.134, No.4, pp.3029.
[8] Narayanan, A., Wang, D. L. (2014). Investigation of Speech Separation as a Front-End for Noise Robust Speech Recognition, IEEE/ACM Transactions on Audio Speech & Language Processing, Vol.22, No.4, pp.826-835.