The application of PLSA features in the automatic assessment system for English oral test

Ding Ming[1*], Dong Bin[1], Yan Yonghong[1], Ding Yousheng[2]

[1]*The Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, No. 21 North 4th Ring Road, Haidian District, 100190 Beijing, China*

[2]*College of Light Industry, Wuhan Polytechnic, No.463 Guan Shan Road, Hongshan District, 430000 Wuhan, China,*

Received 23 November 2014, www.cmnt.lv

Abstract

As an efficacious statistical tool for the analysis of co-occurrence data, the PLSA (Probabilistic Latent Semantic Analysis) is usually applied to the information retrieval. However, the theory foundation of PLSA is document data mining. So PLSA should also be a content understanding tool. In this paper, we try to develop its potential as an content assessment feature extraction tool for the auto English oral test rating system which need more precision and comprehensive content assessment. In the contrast group, word frequency which is extracted from the test data is used to assess the content correlation in the A&Q item as a data mining feature and it has proved to be a success .But, the word frequency feature has a significant weak point: When the system lacks test data, the capability of the feature will drop sharply. Oppositely, building the PLSA model of word frequency with the data prepared before the exam and extracting the probabilistic feature from the examinee's speech can avoid the problem above. In the result, the single dimension feature performance of PLSA feature is better than the simple word frequency feature and the assessment performance will also be improved, if the choice of PLSA model parameters is appropriate.

Keywords: word frequency, PLSA, automatic assessment, oral test

1 Introduction

In China, people and government pay lots of attention on the English education. Recently, more and more people want to take an exam which contains oral test to check their learning effect and capability in communication. For the purpose to satisfy the people's demand, designing an automatic assessment system which can suit the large scale evaluation mission to replace the high costly human rating is necessary. Using computer to assess the examinee's speech has a lot of advantages such as objectivity, consistency and batch processing which cause that the cost of automatic assessment system is much lower than the cost of human rating.

The auto scoring system based on auto speech recognition techniques is widely used in recent twenty years, many research organizations have built their own system. The most of these systems are composed by three main modules-speech recognition module, feature for assessing extraction module and score predicting module [1, 2, 7, 8]. According to the research before, the features' discriminating ability mostly decided the capability of the assessment system. Our system also uses this architecture and finds some suitable features against the particularity of oral A&Q (answer and question) test.

The A&Q item is a kind of test that needs examinee to answer the question according to the clues giving by paper. In the test, the most important evaluation standard is whether the examinee's answer is right or wrong in semantic level. So the posterior probability features which performance well in reading item [3] and the fluency features which performance well in open oral expression item [4] aren't suitable the A&Q item. The content assessment features are the most suitable to be applied to this particular test.

In this paper, we pay our attention on two kinds of content assessment features: one is the word frequency feature; the other is the PLSA feature. Both of them are statistical features which are extracted through the document data mining technique. So they have the capability to discriminate the examinee's answer when we get the enough training data.

The rest of this paper is structured as follows: section 2 introduces the word frequency feature and the way how to extract it; section 3 is a brief introduction of PLSA and its application in rating system; the procedure and result of the experiment is shown in section 4; and finally in section 5 some conclusions will be given.

2 Word frequency feature

Document data mining is usually applied to document retrieval and the document content analysis. There is a keyword-based mining technology which can be used in the assessment system. The core of the technology is the word frequency.

[*] ***Corresponding author's*** e-mail: dingming@hccl.ioa.ac.cn

2.1 WORD FREQUENCY FEATURE

According to the theory, $freq(d,t)$ presents the times of word t occurrence in the document d. In the auto rating system, we define the recognition transcription of one examinee's answer to a question as a document and define all the answers to this question as a kind of document. In this paper, the relative word frequency takes the place of the simple word frequency for reducing the fluency of different kinds of document. The relative word frequency of word w in the same kind of document is expressed as

$$f(w) = \frac{count(w)}{\sum_{i=1}^{N} count(w_i)}, \quad (1)$$

where N is the number of all different kinds of word in all documents that belong to the one type, every word w occurrence in the document will make $count(w)$ add 1. However, the occurrence of the same word in the same document is always counted one time. Each document's word frequency is expressed as

$$P(w_1^s) = \sum_{j=1}^{s} f(w_j), \quad (2)$$

where s is the number of all the word in the document. The word frequency of one document is the base unit in the feature extraction step.

In fact, there are so many English words not having semantic meaning, such as in, at, I, it, but, a, an. These pronouns, prepositions, conjunctions, articles and particles are widely used in the oral test, but they may reduce the accuracy of our system in content assessment. The solution of the problem is setting a list of words called stop words which should be ignored when we calculate the word frequency.

2.2 FEATURE EXTRACTING

The assessment system is combined by three main modules: the speech recognizing module, the feature extracting module, the assessment module. The speaker's speech is the input to the speech recognizing module and the recognition result is the output which is the input to the feature extracting module at the same time. The assessment module is supervised classifier, so we must use the result of the second module and the human score to make a score mapping model. At last the model will give the score according to the assessment feature and mapping rule.

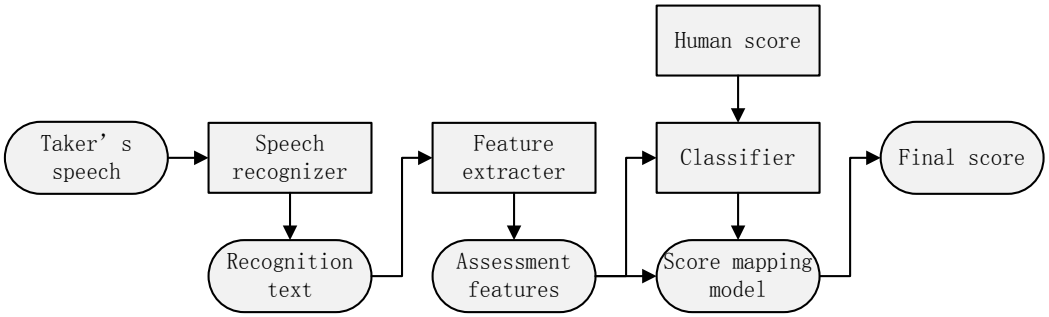

FIGURE 1 Architecture of assessment system for A&Q test.

As the figure 1 show that all the three components of the system are important. However, the key point of the research is the feature extracting module. The word frequency feature belongs to the assessment features and its source is the recognition text. Because the answer speech to the A&Q question is open and various, so the recognition text normally comes from the LVCSR recognizer which is suitable to the large vocabulary and open circumstance recognition. Though extracting the word frequency from the recognition text which is treated as document is not complicated, how to use the word frequency to generate the assessment feature of every examinee is considerable. According to the different way of calculating, we can generate four different kinds of word frequency feature.

First of all, accumulating all the word frequency that belongs to one examinee:

$$L_1(spkr) = \sum_1^m p(w_1^s), \quad (3)$$

where m is the num of document that is equal the num of question need one examinee to answer.

Second is weighted accumulating:

$$L_2(spkr) = \sum_1^m \frac{1}{s} p(w_1^s), \quad (4)$$

The third, all the word frequency composing a vector:

$$\overline{L_3(spkr)} = (\frac{1}{s_1} P\left(w_1^{s_1}\right), \frac{1}{s_2} P\left(w_1^{s_2}\right), \ldots \frac{1}{s_m} P\left(w_1^{s_m}\right)), \quad (5)$$

The fourth is still a vector, but it is different to the third:

$$\overline{L_4(spkr)} = (\sum^{T1} \frac{1}{s} P(w_1^s), \sum^{T2} \frac{1}{s} P(w_1^s), \ldots \sum^{Tl} \frac{1}{s} P(w_1^s)), \quad (6)$$

According to the difference of the question's semantic circumstance, we define different type of question. The word frequency of the same type is accumulated and the values of different types compose the feature vector. This is a human-operated method. Where l is the num of different types.

For the purpose to measure the performance of the four different assessment features, the human score which is

given by some experienced English teacher or professional raters should be the assessment standard. Calculating Pearson r correlation of the human score and the word frequency features is a direct way to find which one may bring more improvement to our system.

The table below shows the performance of each feature. The fluency feature of the system is the baseline as the most successful feature in the open oral test.

TABLE 1 Performance of features

Feature Type	Feature Name	Correlation
Fluency	Speed Pacing	0.508
	Pause Num	0.276
Word Frequency	$L_1(spkr)$	0.503
	$L_2(spkr)$	0.422
	$\overline{L_3(spkr)}$	0.327
	$\overline{L_4(spkr)}$	**0.554**

When the feature is a vector, the value is the average of single dimension feature's correlation.

The result states that the best way to use word frequency feature is extracting feature according to the difference of the question's semantic circumstance. This information should be gotten from the description of the test. For example, the test is usually composed by some parts so that one part should be defined as one type.

Though the frequency is a well-done feature in most case, its performance will sharply drop in some special situation. When the speech data of examinees is not enough, the frequency feature can't work well as a data-based statistical feature. In extreme circumstances such as there is only one examinee to take the test, the frequency may lose the assessment capability. So we have to find a new assessment feature which can adjust to these extreme circumstances.

3 PLSA feature

PLSA is a statistical tool for the analysis of two-mode and co-occurrence data which is well-done in information retrieval, natural language and machine learning from text [9].The automatic assessment system may get some help from its capability in machine learning from text. If the PLSA tool can get some useful information from the speech recognition text, we believe that the performance of our assessment system will be improved.

3.1 PLSA THEORY

In fact, the PLSA is a statistical model which is associated by the probability distribution of three basic parameters. They are document *d*, word *w* and an unobserved parameter *z* which can be called latent semantic variable. A joint probability model is defined by the mixture:

$$P(d,w) = P(d)P(w \mid d),$$
$$P(w \mid d) = \sum_{z \in Z} P(w \mid z)P(z \mid d), \quad (7)$$

It means that *d* and *w* are independent conditional on the state of the associated latent variable. Since the cardinality of *z* is smaller than the num of documents/ words in the collection, *z* acts as a bottleneck variable in predicting. The model can be equivalently parameterized by:

$$P(d,w) = \sum_{z \in Z} P(z)P(d \mid z)P(w \mid z), \quad (8)$$

Two kinds of tasks, predicting words' meaning or predicting documents' category, can be described by this equation. The figure 2 shows these two kinds of tasks.

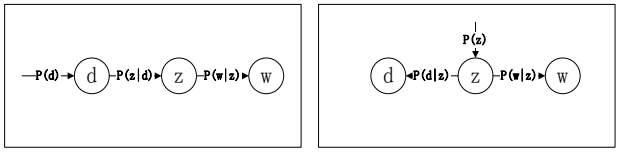

FIGURE 2 PLSA model in the asymmetric and symmetric.

Like most of statistical latent variable models the PLSA model's parameters can be estimated by the Expectation Maximization algorithm [5, 6]. To the PLSA model the E-step equation should be:

$$P(z \mid d,w) = \frac{P(z)P(d \mid z)P(w \mid z)}{\sum_{z' \in Z} P(z')P(d \mid z')P(w \mid z')}, \quad (9)$$

However, the parameters estimating in the M-step should be:

$$P(w \mid z) \propto \sum_{d \in D} n(d,w)P(z \mid d,w),$$
$$P(d \mid z) \propto \sum_{w \in W} n(d,w)P(z \mid d,w), \quad (10)$$
$$P(z) \propto \sum_{d \in D} \sum_{w \in W} n(d,w)P(z \mid d,w),$$

where $n(d,w)$ is the word frequency of the word w in the document d.

3.2 FEATURE EXTRACTION

The task of PLSA in the automatic assessment is predicting documents' category. So the conditional probability $P(z \mid d)$ is our destination. According to the probability formula, $P(z \mid d)$ can be gotten by:

$$P(z \mid d) = \frac{P(z)P(d \mid z)}{\sum_{z \in Z} P(d \mid z)P(z)}, \quad (11)$$

$P(z)$, $P(d \mid z)$ are the basic parameters of the PLSA model. So the first step of extracting assessment features is building a suitable model. According to the test, we prepare some data which come from English experts or excellent students for model building. Then we train the model to get the parameters. As a latent parameter, z can't be observed. So we have to set the value of z in the model initializing step. In the beginning, we set the dimensions of z equaling the num of test item and the value being the inverse of the test items' num.

The second step is renewing the parameters $P(z)$, $P(d \mid z)$ according to the examinee's speech. In the renewing step, the value of $P(w \mid z)$ which is gotten from the training step should be fixed. Usually, the meaning of words in a test will not change anymore.

The third step is calculating the $P(z \mid d)$. Then the value is the assessment feature in the system.

The table below shows the performance of PLSA feature. In the aspect of Pearson r correlation, the new feature is better than old features.

TABLE 2 Performance of features

Feature Type	Feature Name	Correlation
Fluency	Speed Pacing	0.508
	Pause Num	0.276
Word Frequency	$\overline{L_4(spkr)}$	0.554
PLSA	$P(z \mid d)$	**0.623**

Even if there is only one examinee to take the test, the PLSA feature can work well, because the PLSA model can be built without the examinee's speech.

In the follow-up experiments, we find that the initialization of *z* is very important. Our setting in the beginning is not the best choice. In fact, the num of the test item isn't just equal the num of latent semantic category. The performance of the PLSA feature when the *z* has different values will be given in the result of experiments.

4 Experiment

4.1 DATA OF EXPERIMENT

First of all, we built a dataset which was comprised by three parts, the training data, the testing data and the data for PLSA model building. The training data and testing data were from a real A&Q exam, we got the speech data of examinees and their final scores which were given by professional raters. The exam had 16 items. The data distribution of the training data which contained 1000 students' speech was uniform (the data of each item was nearly equal). The data distribution of the testing data which contained 3000 students' speech was close to the real data distribution. The third part had only 200 people's speech, but they were clearer and more precise. This could ensure the PLSA model had assessment capability.

The human score was the integer from 0 to 5. 0 was the worst and 5 was the best. The data of each score level was still uniform. We used the information to build the score mapping model after extracting all the features.

4.2 RESULT OF EXPERIMENT

For the purpose to measure the performance of each feature in assessment task, we set a baseline system which only contained old fluency and pronunciation quality features. Then the word frequency feature and the PLSA feature were added into the system respectively.

In order to get assessment performance directly, we used three main evaluation parameters. They were the expectation of score deviation between human scores and machine scores, the re-examination rate (the rate of whose score deviation is bigger 1) and the correlation coefficient between human scores and machine scores. In one hand, we could compare the performance of baseline system, system with word frequency feature and system with PLSA feature. In the other hand, we could also analyze the influence to the system with different values of z. The result of the experiment was showed in Table 3.

The PLSA feature was better than word frequency feature when we compared performance of single feature, but the performance superiority was not obvious when they were adding into the baseline system. If we can choose an appropriate value of z, we will get better result in assessment.

TABLE 3 Performance of combining all the features

Evaluation Items	**Feature Type Adding into System**				
	None	***Word Frequency Feature***	***PLSA Feature***		
			z=16	***z=10***	***z=5***
Score deviation	0.528	0.499	0.472	**0.470**	0.492
Re-examination rate	0.031	0.029	0.029	**0.026**	0.028
Correlation coefficient	0.642	0.667	0.660	**0.674**	0.658

5 Conclusion

This paper presents our work on how to apply the PLSA tool in the automatic assessment system and shows the advantage of PLSA feature which is better than old features and word frequency feature in system as a single feature. At the same time, the result of experiment states that the application of the PLSA feature improved the performance of assessment system, but it still had more potential. So our future research will focus on how to build a more efficiency PLSA model and improve the way of feature extracting in order to get more advantage.

References

[1] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronuciation scoring of foreign language student speech", *Proc. of ICSLP 96*, pp.1457-1460, Philadelphia, Pennsylvania 1996.

[2] K. Tatsuya, D. Masatake, etc, "Practical use of English pronunciation system for Janpanese students in the CALL classroom", *INTERSPEECH-2004*, pp. 1689-1692, 2004.

[3] H. Franco, H. Bratt, R. Rossier, etc, "A speech recognition and pronumciation scoring toolkit for computer-aided language learning applications", *Language testing*, 2010.

[4] K. F. Lee, S. Hayamizu, and H. W. Hon, "Allophone clustering for continuous speech recognition", *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol.2, pp.749-752, 1990.

[5] R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system", *Proc. ICASSP 90*, *IEEE CS Press*, Piscataway, N. J., pp.129-132, 1990.
[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood with incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, Series B 39, pp.1-38, 1977.
[7] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronuciation scoring of language instruction", *Proc. Int'l Conf. on Acoust, Speech and Signal Processing*, pp.1471-1474, Munich, 1997.
[8] SM Witt, SJ Young, "Phone-level pronunciation scoring and assessment for interactive language learning", *Speech communication*, volume 30, 2000.
[9] T.Hofmann, "Probability Latent Secmantic Analysis", *in the 15th Conference University in AI*, 1999.

Authors

Ding Ming, 1986.11.17,Hubei, P. R. China

Current position, grades: B.Sci degree in acoustics from Nanjing University, Nanjing, P. R. China, in 2009.
University studies: Nanjing University, Nanjing, P. R. China
Scientific interest: English oral automatic assessment
Experience: graduate work in CALL at the Institute of Acoustics in Chinese Academy of Sciences until now.