Chen, W. et al. (Eds.) (2016). Proceedings of the 24th International Conference on Computers in Education. India: Asia-Pacific Society for Computers in Education

Automatic Scoring of English Speaking Test Using Automatic Speech Recognition

Keiji YASUDA*, Hiroyuki KAWASHIMA*, & Hiroaki Kimura*
**KDDI R&D Laboratories, Inc.*
{ke-yasuda,hi-kawashima,ha-kimura}@kddilabs.jp

Abstract: In this paper, we propose an automatic scoring method for conversational English test using automatic speech recognition and machine learning techniques. We administered a mock English speaking test with 111 Japanese speakers. According to the experimental results using the data, the correlation between human expert scoring results and automatic scoring results was 0.825.

Keywords: English speaking test, automatic evaluation, machine learning, SVR

1. Introduction

English Communicative skills are becoming more important due to globalization of economic activities, widespread use of the Internet, and other factors. There are four language skills: reading, listening, speaking, and writing (CEFR, 2001). Measuring learners' proficiencies in these skills is necessary in the language learning phase. Computer Based Testing (CBT) offers a short turnaround time with the potential to boost learning efficiency.

Here, we discuss the relationship between the four language skills and the testing method. Listening and reading skills are suitable for conventional multiple choice testing. Hence, these skills are easy to measure with CBT. On the other hand, writing and speaking skills take ingenuity to measure with CBT because conventional multiple choice testing is inappropriate to measure proficiency in these skills. For writing skill, an automatic essay scoring method based on text coherence has been proposed by Crossley and McNamara (2011). For speaking skill, the Pearson Versant system was developed from the original Phone Pass ASR-based test by Ordinate. It tests vocabulary and fluency in addition to pronunciation and word ordering (Pearson, 2011).

In this paper, we propose an automatic method for a high spontaneity speaking test using Automatic Speech Recognition (ASR) and machine learning techniques. Compared with the conventional speaking test service, this method is more suitable for measuring a practical speaking skill in a concrete topic or field as well as evaluation using can-do descriptors, such as Common European Framework of Reference for Languages (CEFR, 2011). Although the experiments in this paper were carried out using Japanese non-native English speakers, the proposed method is highly data driven method and is portable to other mother tongues and other languages as far as the data are available.

Section 2 explains the English speaking test. Section 3 describes the proposed automatic scoring system. Section 4 shows the evaluation experiments and results. Finally, section 5 concludes the paper.

2. English Speaking Test

Here, we explain the English speaking test material. In this experiment, we used an English speaking test from one of the largest online English lesson providers, the Rare Job Company. The test consisted of six sections. Each section was bounded by time up to two minutes. Total testing time for one examinee was about 30 minutes. The test measured six English skills—Ability to Express (AE), Ability to Make sentences (AM), Ability to Understand (AU), Pronunciation (P), Grammar (G), and Fluency (F). The relationship between each section and measured English skills are shown in table 1.

3. Developed Automatic Scoring System

3.1 Non-native Automatic Speech Recognition

We used an Automatic Speech Recognition (ASR) system adapted to Japanese non-native English speakers and a speaking test. The speech data for the adaptation was collected by mock speaking test using the previously mentioned test material. The speech and the transcription were used for Deep Neural Net (DNN) acoustic model adaptation and *n*-gram language model adaptation. By using the acoustic model and language model adaptation, word accuracy improved to 62.3% from baseline performance of 29.8%.

Table 1: Details of the test (Relationship between sections and evaluation criteria)

Section	Test type	Evaluation Criteria					
		AE	AM	AU	F	G	P
1	Self-Introduction	✔	✔	✔	✔	✔	✔
2	Reading			✔	✔		✔
3	Answering Questions			✔	✔	✔	✔
4	Role play	✔	✔	✔	✔	✔	✔
5	Describing a picture	✔	✔		✔	✔	✔
6	Summarization	✔	✔	✔	✔	✔	✔

(AE: Ability to express, AM: Ability to make sentences, AU: Ability to understand

P: Pronunciation, G: Grammar, F: Fluency)

3.2 Automatic Scoring Method

Figure 1 shows the model training phase of the proposed system. First, speech in the speaking test was manually scored by human experts. Second, the same speech was recognized by ASR. Then, the system extracted the linguistic features from the ASR output. Details of the features are explained in section 4. Finally, the extracted features and scoring results by human experts were input to machine learning to build an automatic scoring model. For the actual scoring phase, the system automatically scored the examinees' speech using the extracted features of the ASR results and the automatic scoring model.

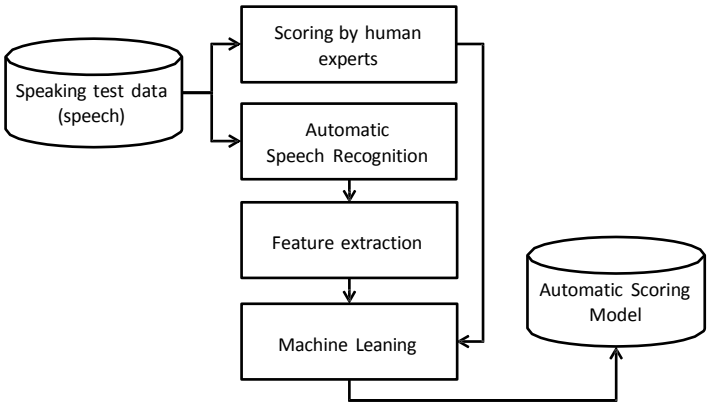

Figure 1. Proposed method (Training of automatic scoring model)

4. Experiments

4.1 Experimental Setting

For data collection, we carried out a mock speaking test with 111 non-native English speakers. All of the human subjects spoke Japanese as the mother tongue. Before the data collection, we sorted human subjects by TOEIC scores so as to collect uniform distribution data of English skills. For machine learning, we used the Support Vector Regression (Cortes and Vapnik, 1995). And the following three features were used for model training.

- ✓ Bag of words of ASR results
- ✓ Number of words uttered by each learner
- ✓ Vocabulary size of each examinee's utterance

4.2 Experimental Results

For evaluation of automatic scoring, we introduced two evaluation measures. The first measure was the Root Mean Square Error (RMSE) from scoring results by human experts. The second measure was the correlation between scoring results by human experts and automatic scoring results. We averaged the scoring results of three to four human experts, and then the average score was used as an oracle score for model training and evaluation. Table 2 shows the evaluation results of automatic scoring using the 10-fold cross validation test on the 111 mock exam test results. We trained the model for each of the six evaluation criteria excluding the overall score. Then, six scores were summed to determine the overall score.

As shown in table2, the estimation of the Overall Score (OS) has a 0.825 correlation coefficient, which is the highest value out of the seven criteria, that is, the estimated overall score is the most reliable estimated score out of the seven evaluation criteria. Meanwhile, the estimated scores of Ability to express and Pronunciation have low correlation coefficients which are less than 0.8. The reason for the low correlations is that acoustical features are not used for machine learning.

Table 2. Evaluation results.

	Evaluation Criteria						
	AE	AM	AU	F	G	P	OS
Full Score	80	40	50	60	50	60	340
Correlation	0.757	0.806	0.807	0.813	0.822	0.786	**0.825**
RMSE	5.11	2.83	3.94	4.307	3.400	4.372	21.784

(AE: Ability to express, AM: Ability to make sentences, AU: Ability to understand, P: Pronunciation, G: Grammar, F: Fluency, OS: Overall Score)

5. Conclusions and Future Work

We proposed an automatic scoring method for a conversational English speaking test using ASR and machine learning techniques. We carried out experiments using test results from 111 Japanese non-native speakers of English. According to the experimental results, the proposed method showed satisfactory performance for the overall score. The correlation and RMSE between human expert results and automatic scoring results are 0.825 and 21.784, respectively.

As future work, we will try to add a new feature, such as acoustic likelihood for machine learning, to improve scoring performance of partial scores as related to pronunciation.

Acknowledgements

We wish to thank to Rare Job for providing test materials and evaluations by human experts.

References

Cortes C. and Vapnik V. (1995). Support-vector network. *Machine Learning*, 20, 273-297.

Council of Europe (2001). Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge: *Cambridge University Press*.

Crossley, S. A. & McNamara, D. S. (2011). Text Coherence and Judgments of Essay Quality: Models of Quality and Coherence, *Proc. of the 29th Annual Conference of the Cognitive Science Society*, 1236-1241.

Pearson (2011), Versant English –Test Description and Validation Summary–, http://www.versanttest.com/technology/VersantEnglishTestValidation.pdf