Open Answer Scoring for S-CAT Automated Speaking Test System Using Support Vector Regression

Yutaka Ono*, Misuzu Otake*, Takahiro Shinozaki*, Ryuichi Nisimura†,
Takeshi Yamada‡, Kenkichi Ishizuka‡, Yasuo Horiuchi*, Shingo Kuroiwa* and Shingo Imai‡
* Chiba University, Chiba, Japan http://www.ailab.tj.chiba-u.jp
† Wakayama University, Wakayama, Japan http://www.wakayama-u.ac.jp/~nisimura/
‡ University of Tsukuba, Ibaraki, Japan http://www.tsukuba.ac.jp

***Abstract*—We are developing S-CAT computer test system that will be the first automated adaptive speaking test for Japanese. The speaking ability of examinees is scored using speech processing techniques without human raters. By using computers for the scoring, it is possible to largely reduce the scoring cost and provide a convenient means for language learners to evaluate their learning status. While the S-CAT test has several categories of question items, open answer question is technically the most challenging one since examinees freely talk about a given topic or argue something for a given material. For this problem, we proposed to use support vector regression (SVR) with various features. Some of the features rely on speech recognition hypothesis and others do not. SVR is more robust than multiple regression and the best result was obtained when 390 dimensional features that combine everything were used. The correlation coefficients between human rated and SVR estimated scores were 0.878, 0.847, 0.853, and 0.872 for fluency, accuracy, content, and richness measures, respectively.**

I. Introduction

As the globalization proceeds, the number of international students is increasing and there are demands for evaluating their language ability for various purposes. Along with tests with admission fee [1], [2], Japanese Computer Adaptive Test (J-CAT) [3] had been developed as a free online proficiency test for Japanese language learners. It is a computerized adaptive testing based on item response theory [4] that adapts to examinee's ability level so that minimum number of question items is required to estimate the examinee's ability [5]. The J-CAT test consists of four sections that are respectively designed to examine vocabulary, grammar, reading, and listening related abilities. The testing system consists of an item bank and a testing engine, and is fully automated. The score report is immediately provided to the examinee at the end of the test. The first pre-test was conducted in 2007 and it has been used by 26 institutions around the world since then. The system is used around 5000 examinees per year.

Despite the success, a limitation of J-CAT is the lack of speaking tests while speaking ability is very important for international students in their daily lives. For this reason, S-CAT is under development that extends J-CAT by supporting speaking test items. It will be the first automated adaptive speaking test for Japanese language. Since the test is provided for free, it is not possible to score the spoken answers by human raters. Therefore, speech processing techniques are employed. Using computers for scoring has also a benefit that it is free from biases by individual human raters.

The S-CAT test will have five item categories: reading, multiple-choice, blank-filling, sentence generation, and open answer. The reading is to read a given sentence, the multiple-choice is to select an answer from choices and to read it, the blank-filling is to pronounce words that best fit in a blank. The sentence generation is to compose a sentence that fits to a question and to speak it, and open answer is to argue about a given topic. While the automatic scoring is required for all the categories, the focus of this paper is the open answer. It is technically the most difficult because the answer utterances have the largest freedom.

Our basic strategy for this challenge is to first prepare a pair of data consisting of waveforms of spoken answers by examinees and their scores rated by human raters. Then we extract various features from the waveforms and train score estimators so that the human scores are predicted based on the features. Related research has been performed by Educational Testing Service (ETS) for English where a multiple regression was used to score spontaneously spoken responses from examinees based on five and 11 features [6]. They reported correlations from 0.57 to 0.68 between the human rated and regression estimated scores. One difference of our research from it is simply the target is Japanese instead of English, but another difference is the utilization of support vector regression (SVR) which is expected to be more robust than multiple regression for overtraining. This property of SVR have motivated us to use large number of features. In fact, we investigate to use nearly 400 features.

The organization of the rest of this paper is as follows. In Section II, the overview of the S-CAT system is described. In Section III, the open answer items in S-CAT are explained. In Section IV, the principles of SVR is briefly reviewed. Features we used for the answer scoring are explained in Section V. Experimental conditions are described in Section VI and the results are shown in Section VII. Finally, conclusions are given in Section VIII.

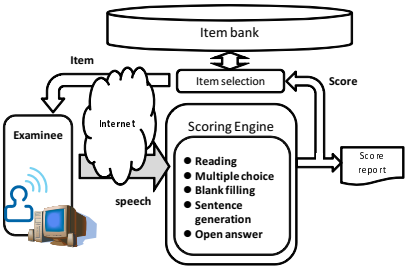

Fig. 1. S-CAT system overview.

II. System Overview

Figure 1 shows the overview of the S-CAT system. The system is based on the client/server model. Clients and the S-CAT server are connected by the HTTP protocol and items are presented to examinees via web browsers. When an examinee answers an item by voice, it is recorded by a microphone and transmitted to the S-CAT server over the Internet. The scoring engine scores the answer and the next item is selected based on the score so that the examinee's ability level is efficiently identified.

The scoring engine is a set of score estimators designed for each item or for each item category. As it has been mentioned in the introduction, the focus of this paper is the automatic scoring of the open answer items. We use SVR as an estimator with various features.

III. Open answer items in S-CAT

The open answer question items in S-CAT are items that ask examinees to talk about their thinkings or to explain something that can be read from a given graphic chart or from a leaflet. The questions are given by Japanese voice but they are something like this: "Are you for solar energy or not? Why?" Another example is a combination of a voice instruction and a picture of a leaflet: "Assume you have gotten this leaflet today. Call your friend and explain what kind of event is going to be held and where it is. Note however, your friend is absent and you have to record your voice to an answering machine." The examinees are required to talk for 40 seconds per item.

Answers from examinees are scored by the following four measures.

Fluency:
: The fluency of the pronunciation.

Accuracy:
: The correctness of the syntax and the relevance of the wordings.

Contents:
: The degree of how the requested task is accomplished. Whether the necessary information is transmitted correctly.

Richness:
: Abundance of vocabulary and expression.

For the training data used for system development, these scores are rated by at least three human specialists per item and their ratings are averaged. The scores rated by each specialist are integers from zero to four where four being the best. The averaged scores are real numbers.

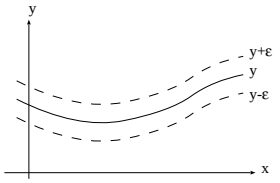

Fig. 2. SVR regression curve and ϵ-insensitive tube.

The human raters rate the spoken answers comprehensively using their full knowledge. However, we do not imitate their rating process for the automatic scoring. Instead, we use machine learning techniques to predict human rated scores based on features that are automatically extracted from speech waveforms.

IV. Support Vector Regression (SVR)

For binary and multi-class classification problems, Support Vector Machine (SVM) is widely used. It is trained so as to maximize margin of the decision boundary, which results in a sparse solution and makes SVM robust for overtraining. SVR is an extension of SVM for regression problems [7]. It predicts continuous target values based on Equation shown in (1) using a kernel function k.

$$y(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n k(\mathbf{x}, \mathbf{x}_n) + \beta. \quad (1)$$

In the equation, N is the number of training samples, $\mathbf{x}_n$ is feature vector, α_n and β are coefficients. The cost function used for training SVR gives zero error if the absolute difference between the prediction and the target is less than ϵ for a positive value $\epsilon > 0$ as shown in Figure 2, which makes SVR depends only on a subset of the training data called support vectors. The coefficients α_n for points within the ϵ-insensitive tube are all zero and they do not contribute to the prediction. SVR is similar to the multiple linear regression but the difference is this sparsity of the solution.

V. Features for open answer scoring

For the score estimation, we use decoder based features that are based on speech decoder outputs and acoustic features that are extracted from waveforms without using speech decoders. The decoded features can capture what are spoken, but they are affected by recognition errors. On the other hand, acoustic features are free from the decoding errors but cannot capture what are spoken.

The decoded features we investigated are as follows. Decoders are used to recognize what words exist in speech waveform and how they are time aligned.

VOA:
: A measure for abundance of vocabulary defined as $\frac{W_{unq}}{\sqrt{2W_{tot}}}$, where W_{unq} is the number of unique words and W_{tot} is the total number of words [8].

ROS:
: Two rate of speech measures are defined. One is the

number of syllables divided by the length of speech segments. The other is the length of speech segments divided by the recording length.

KWD:
Number of keywords found in answer speech. Around 80 keywords are manually defined for each item that closely related to the topic, and the umber of these keywords found in recognition hypothesis is counted. Since it is difficult to accurately recognize utterances given by international students with varying language skills, two different decoders are used simultaneously and the number of keywords are counted for each recognition hypothesis. Additionally, the number of keywords found in one of or both of the hypotheses are also used as features. The total dimension of the feature is four.

DCDs:
Combined features of VOA, ROS, and KWD.

Five types of acoustic features listed in the followings are used. To extract these, first a frame based features and their deltas are estimated. Then, their 12 kinds of statistics are computed as the features of a speech segment. These statistics are minimum, maximum, range, positions of minimum and maximum, arithmetic mean, slope, offset, quadratic error, standard deviation, skewness and kurtosis.

ENG:
The statistics of signal frame energy.

ZCR:
The statistics zero-crossing rate.

VPR:
The statistics of voicing probability.

F0:
The statistics of fundamental frequency. Among the 12 statistics, minimum was not used.

MFC:
The statistics of twelve dimensional Mel-frequency cepstral coefficients (MFCC).

ACTs:
Combined features of ENG, ZCR, VPR, F0, and MFC.

These are extracted using the openSMILE feature extractor with the default configuration setting for the INTERSPEECH 2009 Emotion Challenge feature set [9]. Please refer to the openSMILE manual for the details. However, a difference is that minimum of F0 has been removed since it is always zero for our data. Hence, the total dimension is 383. Before they are used for the score estimations, each of the feature dimensions is normalized by a mean and a standard deviation estimated on a training set.

VI. Experimental setup

In order to collect data for the development of S-CAT, we made a web based system that imitated the actual speaking test. For the experiments reported here, samples from 101 subjects were used that have been collected so far using the web system. These subjects were international students and had varying Japanese skills. The number of open answer items was 10 and each of the subjects answered all these questions providing 1010 samples in total. For the system development and evaluation, the data was divided to training and evaluation sets. Samples from 81 subjects were used as the training set and the ones from 20 subjects were used as the evaluation set. Since S-CAT is designed to be operated using a predefied item bank, the experiments were performed subject-independent and item-closed condition. The averaged variance between human raters in each item for the test set was 0.355.

To extract the decoder based features, we used the Julius [10] and the T^3 [11] decoders. The acoustic model was a triphone HMM trained on the CSJ Japanese corpus [12] and adapted to the utterances in the training set. The language model was a tri-gram trained using the transcribed text of the training data and texts from the web and news articles that amounted to 300M words in total. Considering the difficulty of the task, we focused on word correctness rather than accuracy so as to minimize the loss of information by deletion errors, and it was 60.6% for julius and 73.2% for T3 for the test set. A single estimator is trained and used for all the 10 items in common. SVR is trained and evaluated with SVM-Light [13] using linear kernel and the default parameter setup based on some preliminary experiments.

TABLE I
Fluency score estimation using decoded features

Features		Regression		SVR	
Type	dim	Correl	RMS	Correl	RMS
VOA	1	0.861	0.787	0.861	0.681
ROS	2	0.818	0.806	0.820	0.829
KWD	4	0.728	1.034	0.703	1.161
DCDs	7	0.854	0.719	0.863	0.850

TABLE II
Fluency score estimation using acoustic features

Features		Regression		SVR	
Type	dim	Correl	RMS	Correl	RMS
ENG	24	0.765	0.851	0.709	0.942
ZCR	24	0.820	0.786	0.760	0.933
VPR	24	0.783	0.903	0.629	1.194
F0	23	0.705	1.056	0.684	1.233
MFC	288	0.854	0.711	0.854	0.723
ACTs	383	0.773	0.862	0.865	0.683

VII. Experimental Results

Table I shows Pearson's correlation and root mean square error (RMS) of the estimated fluency scores by multiple linear regression and SVR. Among the three types of baseform decoded features (i.e. VOA, ROS, and KWD), VOA gave the best results. The largest correlation 0.863 was obtained when all the features are combined (i.e. DCDs) and SVR was used. Although, the lowest RMS 0.681 was obtained when VOA was used with SVR.

Table II shows the results when the acoustic features were used. Among the five types of baseform acoustic features, MFCC gave the best performance. It can also be seen that while multiple regression worked better than SVR when the dimension of the features were small, SVR outperformed it

TABLE III
SCORE ESTIMATION USING DECODED AND ACOUSTIC FEATURES

Features		Estimator	Fluency		Accuracy		Content		Richness	
Type	dim		Correl	RMS	Correl	RMS	Correl	RMS	Correl	RMS
DCDs	7	REG	0.854	0.719	0.817	0.772	0.823	0.794	0.839	0.755
		SVR	0.863	0.850	0.829	0.889	0.834	0.825	0.852	0.751
ACTs	383	REG	0.777	0.860	0.758	0.844	0.794	0.832	0.825	0.764
		SVR	0.865	0.683	0.829	0.711	0.844	0.731	0.860	0.681
VOA+ACTs	384	REG	0.783	0.849	0.762	0.835	0.796	0.830	0.832	0.748
		SVR	0.871	0.676	0.832	0.707	0.845	0.731	0.864	0.673
DCDs+ACTs	390	REG	0.814	0.795	0.802	0.775	0.818	0.797	0.848	0.724
		SVR	0.878	0.654	0.847	0.674	0.853	0.716	0.872	0.651

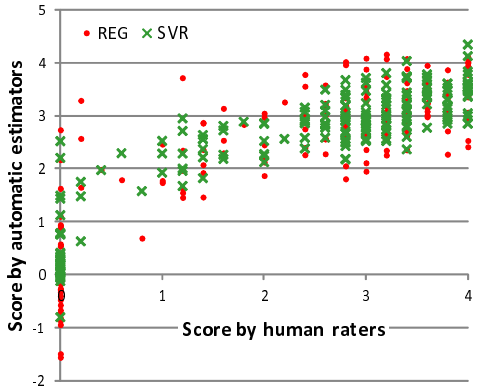

Fig. 3. Scatter plot of human and automatic scores for fluency score.

when the dimension was large. This shows the robustness of SVR for overtraining and its ability to utilize information contained in high dimensional features. The highest correlation of 0.865 and the lowest RMS of 0.683 were obtained by SVR using all the features in combination. These are comparable as the best ones obtained by the decoded features.

Figure 3 shows scatter plot when DCDs and ACTs were used in combination, to estimate the fluency score. It can be seen that SVR results have narrower distribution than multiple regression results.

Table III shows the results of estimation for all the four types of rating scores; fluency, accuracy, content, and richness. In addition to DCDs, ACTs, and their combination (DCDs+ACTs), a combination of VOA and ACTs was also evaluated as VOA was specially good among the decoded features. As can be seen, acoustic features gave generally good performance as decoded features for all the types of rating scores despite they do not aware the contents of utterances. This is probably because the four types of ability of examinees are correlated each other. The best results were obtained when all the features were combined and SVR was used. The correlation for fluency, accuracy, content, and richness were 0.878, 0.847, 0.853, and 0.872, respectively.

VIII. CONCLUSIONS

In order to automatically score spoken answers for open question test items, SVR based estimators were investigated using various decoded and acoustic features. It has been shown that SVR is robust for overtraining and allows the use of high dimensional features while the performance of multiple regression degrades. Acoustic features worked well as decoded features, and the best results were obtained by SVR using all the features in combination. The correlations between human rated scores and the estimated ones were 0.878, 0.847, 0.853, and 0.872, respectively, for the fluency, accuracy, content, and richness measures.

ACKNOWLEDGMENTS

This research has been supported by KAKENHI (22242014). Part of this research has also been supported by KAKENHI (21300066).

REFERENCES

[1] Japan Student Services Organization, Ed., *Examination for Japanese University Admission for International Students*, Bonjinsha Inc., 2011.
[2] Japan Foundation and Japan Educational Exchanges and Services, Eds., *New Japanese-Language Proficiency Test Guidebook: An Executive Summary and Sample Questions for N1, N2 and N3*, Bonjinsha Inc., 2009.
[3] Shingo Imai, Sukero Ito, Yoichi Nakamura, Kenichi Kikuchi, Yayoi Akagi, Hiromi Nakasono, Akiko Honda, and Takekatsu Hiramura, "Features of J-CAT (Japanese computerized adaptive test)," in *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*, 2009, pp. 1–8.
[4] F. M. Lord, *Applications of Item Response Theory To Practical Testing Problems*, Routledge, 1980.
[5] D. J. Weiss and G. G. Kingsbury, "Application of computerized adaptive testing to educational problems," *Journal of Educational Measurement*, vol. 21, pp. 361–375, 1984.
[6] Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883 – 895, 2009.
[7] H. Drucker, C. J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *NIPS*, 1996, pp. 155–161.
[8] M. Tajima, A. Fukada, and N. Sato, "A study of the validity of the indices expressing lexical variation : Using a corpus of l1 Japanese writing by college-level students," in *Proceedings of the Research Institute of Social Systems, Chuo Gakuin University*, 2008, vol. 9(1), pp. 51–62.
[9] Florian Eyben, Martin Wöllmer, and Björn Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM)*, 2010, pp. 1459–1462.
[10] A. Lee, T. Kawahara, and S. Doshita, "An efficient two-pass search algorithm using word trellis index," in *Proc. ICSLP*, 1998, pp. 1831–1834.
[11] P. R. Dixon, D. A. Caseiro, T. Oonishi, and S. Furui, "The titech large vocabulary wfst speech recognition system," in *Proc. IEEE ASRU*, 2007, pp. 443–448.
[12] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," in *Proc. SSPR2003*, 2003, pp. 135–138.
[13] T. Joachims, *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*, MIT Press, 1999.