AUTOMATIC SCORING METHOD CONSIDERING QUALITY AND CONTENT OF SPEECH FOR SCAT JAPANESE SPEAKING TEST

Naoko Okubo[1†], *Yuto Yamahata*[1], *Takeshi Yamada*[1], *Shingo Imai*[1], *Kenkichi Ishizuka*[1], *Takahiro Shinozaki*[2], *Ryuichi Nisimura*[3], *Shoji Makino*[1], *Nobuhiko Kitawaki*[1]
[1]University of Tsukuba, Ibaraki, Japan
[2]Chiba University, Chiba, Japan
[3]Wakayama University, Wakayama, Japan
[†] okubo@mmlab.cs.tsukuba.ac.jp

ABSTRACT

We are now developing a Japanese speaking test called SCAT, which is part of J-CAT (Japanese Computerized Adaptive Test), a free online proficiency test for Japanese language learners. In this paper, we focus on the sentence-reading-aloud task and the sentence generation task in SCAT, and propose an automatic scoring method for estimating the overall score of answer speech, which is holistically determined by language teachers according to a rating standard. In that process, teachers carefully consider different factors but do not rate the scores of them. We therefore analyze how each factor contributes to the overall score. The factors are divided into two categories: the quality of speech and the content of speech. The former includes pronunciation and intonation, and the latter representation and vocabulary. We then propose an automatic scoring method based on the analysis. Experimental results confirm that the proposed method gives relatively accurate estimates of the overall score.

Index Terms— J-CAT, SCAT Japanese speaking test, automatic scoring, quality of speech, content of speech

1. INTRODUCTION

J-CAT (Japanese Computerized Adaptive Test) [1, 2] is a free online proficiency test for Japanese language learners. It is an adaptive test based on item response theory [3], which enables to reduce the number of questions and to estimate the proficiency of an examinee precisely. J-CAT consists of four sections respectively designed to evaluate listening comprehension, vocabulary, grammar, and reading comprehension. It was adopted by 26 institutions around the world and taken by about 5000 people last year. To evaluate speaking comprehension in J-CAT, we are now developing a Japanese speaking test called SCAT (Speaking section of J-CAT), which will be a first automated adaptive speaking test for Japanese language learners. It consists of five tasks: sentence-reading-aloud, multiple-choice, blank-filling, sentence generation, and open answer. The technical difficulty of the automatic scoring increases in this order, since it must accept a variety of answers.

In this paper, we focus on the sentence-reading-aloud task and the sentence generation task in SCAT, and propose an automatic scoring method for estimating the overall score of answer speech. The overall score is holistically determined by language teachers according to a rating standard. In that process, teachers carefully consider different factors but do not rate the scores of them. We therefore analyze how each factor contributes to the overall score. The factors are divided into two categories: the quality of speech and the content of speech. The former includes pronunciation and intonation, and the latter representation and vocabulary. Our proposed method is based on the analysis. It first estimates the scores of each factor from features of input speech and then determines the overall score from the estimates. Experimental results confirmed that the proposed method gives relatively accurate estimates of the overall score.

The rest of this paper is organized as follows. Sect. 2 provides the overview of the proposed method. Sect. 3 and Sect. 4 describe the details of the proposed method and verify its effectiveness for the sentence-reading-aloud task and the sentence generation task, respectively. Sect. 5 summarizes the paper.

2. OVERVIEW OF THE PROPOSED METHOD

The overall scores of the sentence-reading-aloud and the sentence generation tasks are rated holistically based on a rating standard by language teachers. In that process, teachers carefully consider different factors, including pronunciation and intonation. The basic idea of the proposed method is to utilize the factors that contribute to the overall score.

Fig. 1 illustrates the overview of the proposed method. First, the proposed method extracts the features, which reflect the scores of each factor, from the input answer speech. The scores of each factor are then estimated separately by using the extracted features. The target factors are selected in advance, for example, pronunciation and fluency were adopted for the sentence-reading-aloud task as described later. Finally, the overall score is determined by substituting the estimates

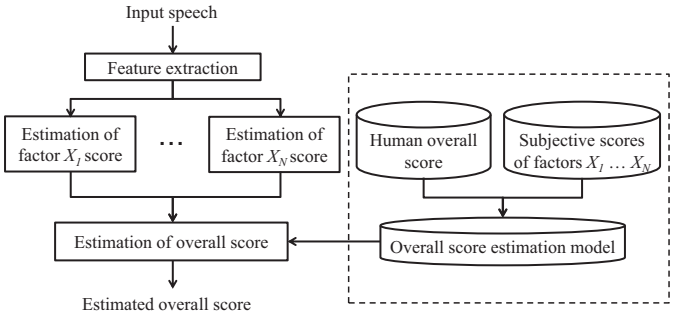

Fig. 1. Overview of the proposed method.

of the scores of each factor in the overall score estimation model. The overall score estimation model is also obtained in advance by using the pairs of the human overall score and the human scores of each factor.

The unique point of the proposed method is to estimate the overall score from the estimated scores of each factor. It leads to the following advantages.

- We can concentrate on estimating the scores of each factor, which requires a limited number of features. This is easier than estimating the overall score directly from many features.
- We can easily find out the features which reflect the scores of each factor. We also effectively add or change the features used for each factor according to the tasks. This facilitates to improve the estimation performance of the overall score.

3. SENTENCE-READING-ALOUD TASK

3.1. Factors to be considered

In the sentence-reading-aloud task, examinees read out aloud a short sentence displayed on computer screen and also presented through headphones. The aim of the task is to evaluate the ability to speak like native Japanese people. It means that the overall score is mainly affected by factors in the quality of speech rather than those in the content of speech.

Referring to the previous works [4] and [5], Pronunciation (X_1), Accent (X_2), Intonation (X_3), Fluency (X_4), and Loudness (X_5) were picked up as candidate factors, which belong to the quality of speech. Note that Fluency (X_4) is defined as smoothness related to time, for example, it may have a low score when speaking haltingly.

3.2. Overall score estimation model

A subjective experiment was conducted to select the factors, which are effective for estimating the overall score, and to obtain the overall score estimation model for the sentence-reading-aloud task.

Table 1. Score rating scale.

Score	Description
4	Excellent (Natural)
3	Good (Unnatural but well understandable)
2	Fair (Unnatural and fairly understandable)
1	Poor (Unnatural and hard to understand)
0	Bad (Not understandable)

Table 2. Correlation coefficients among all the factors in the sentence-reading-aloud task.

	X_1	X_2	X_3	X_4	X_5
X_1	1.00				
X_2	0.82	1.00			
X_3	0.79	**0.92**	1.00		
X_4	0.86	**0.89**	0.86	1.00	
X_5	0.51	0.40	0.41	0.36	1.00

Speech samples were collected by the prototype system of SCAT. Twenty examinees who are Japanese language learners with six different mother tongues answered three questions. Five subjects listen to these 60 speech samples through headphones in a soundproof room and evaluate them. In evaluating one speech sample, the subjects listen to it with focusing on one of the factors, and then rate the score. This is repeated until all the factors are evaluated. The score rating scale used is shown in Table 1. The subjects were instructed to evaluate the speech samples comparing with correct and standard Japanese used at daily conversation. Furthermore, the overall scores of the same speech samples were determined by three language teachers, based on the rating standard with five-level score rating scale. As a result, we obtained the averaged overall score and the averaged scores of each factor per speech sample.

Table 2 shows the correlation coefficients among all the factors. From Table 2, we can see that Intonation (X_2) has a strong correlation with Accent (X_3) and Fluency (X_4). The possible reason is that all the sentences to be read out aloud are short and declarative. By the stepwise selection method based on the linear regression among the overall scores and the scores of each factor, we selected Pronunciation (X_1) and Fluency (X_4) as effective factors and defined the overall score estimation model in Fig. 1 by

$$\text{Overall score} = 0.27X_1 + 0.42X_4 + 1.30. \qquad (1)$$

Since the task contains many questions with a wide range of difficulty, it is desirable to prepare the model specialized in each question. It is however unrealistic to conduct the above experiment every time when a new question is added. We therefore decided to prepare the single model that is applica-

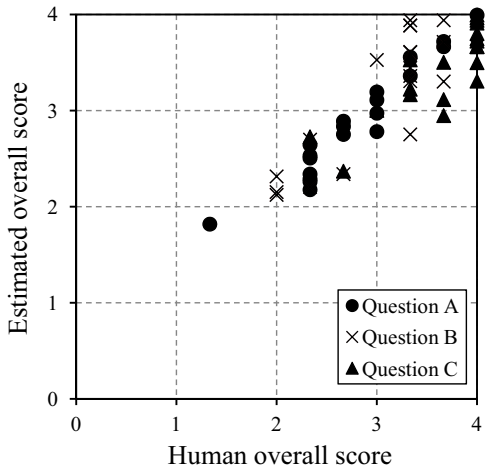

Fig. 2. Relationship between the human overall score and the estimated overall score from the human factor scores in the sentence-reading-aloud task.

ble to different questions.

Fig. 2 shows the relationship between the human overall score and the overall score estimated by substituting the human scores of each factor in Eq. (1). The human scores of each factor are the same as those used for obtaining Eq. (1). In Fig. 2, each point corresponds to one of the speech samples. The correlation coefficient and the RMSE (Root Mean Square Error) are 0.89 and 0.30, respectively. This is the upper limit of the estimation accuracy, which can be achieved when we could estimate the scores of each factor correctly.

3.3. Features to estimate the scores of each factor

To estimate the scores of Pronunciation (X_1) and Fluency (X_4), features based on speech decoder outputs are used. A speech decoder is generally used to recognize what words exist in the input speech and how they are temporally aligned.

We first describe features for Pronunciation (X_1). To evaluate the goodness of pronunciation, conventional methods focus on the acoustic likelihood and its derivatives, which are obtained by using a speech decoder with native speaker acoustic models and/or non-native speaker acoustic models [6, 7, 8]. However, it is well-known that the acoustic likelihood is affected by not only pronunciation but also speaker individuality and a recording condition. To cope with this problem, we propose the following two features, which are calculated by using the time alignment to the sentence to be read out aloud and the time alignment to the output of continuous phoneme recognition without any language limitation.

x_{1a} : Ratio of the number of the frames in which the same phoneme is observed, to the total number of the frames.

x_{1b} : Ratio of the number of the frames in which the different phoneme is observed and the difference between the acoustic likelihoods exceeds a threshold ($\theta = 2.25$), to the total number of the frames.

Note that the features are calculated after all the silence frames are removed. x_{1a} and x_{1b} focus on good pronunciation and especially bad pronunciation, respectively. Hence, x_{1a} would become large when examinees read out aloud a sentence by good pronunciation, opposite to x_{1b}.

We then discuss features for Fluency (X_4). By observing the answer speech samples by the examinees, we found out that Fluency (X_4) can be related to the length of silence in the speech period and the duration time of syllables. Based on the findings, we propose the following two features, which are calculated by using the time alignment to the sentence to be read out aloud.

x_{4a} : Ratio of the number of the silence frames, to the total number of the frames.

x_{4b} : Coefficient of variation of the duration time of syllables.

x_{4a} focuses on whether examinees speak haltingly or not. Note that x_{4a} is calculated after removing the silence frames at the beginning and end of the speech sample. x_{4b} also focuses on the variation of the duration time of syllables, which is based on the fact that we often feel unnatural when the duration time of syllables changes widely. Note that the coefficient of variation is defined by dividing the standard deviation by its average.

3.4. Effectiveness of the Proposed Method

We first examined the effectiveness of the features by estimating the scores of each factor. The speech samples are the same as those used for the subjective test in Sect. 3.2. To calculate the features, the Julius decoder [9] is used with the speaker independent phonetic tied-mixture triphone models [10], which were trained by using native Japanese speech. By the linear regression among the human scores of Pronunciation (X_1) and the values of x_{1a} and x_{1b}, we defined the estimator for Pronunciation (X_1) by

$$X_1 = 1.03x_{1a} - 8.90x_{1b} + 3.05. \quad (2)$$

We also expressed the estimator of Fluency (X_4) by

$$X_4 = -4.77x_{4a} - 2.39x_{4b} + 4.67. \quad (3)$$

We then estimated the scores of each factor to substitute the values of the features in Eq. (2) and Eq. (3), respectively. The estimation accuracy for Pronunciation (X_1) and Fluency (X_4) is summarized in Table 3. We confirmed that a relatively accurate estimate can be obtained.

Finally, we determined the overall score by substituting the scores of each factor estimated above in Eq. (1). Fig. 3 represents the relationship between the human overall score and the estimated overall score. The correlation coefficient

Table 3. Estimation accuracy for X_1 and X_4.

Factor	Correlation coefficient	RMSE
X_1	0.78	0.48
X_4	0.80	0.60

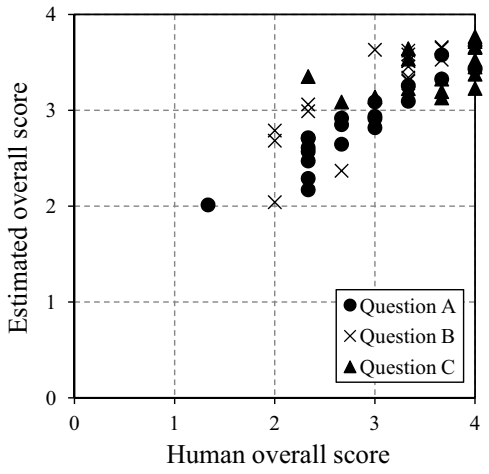

Fig. 3. Relationship between the human overall score and the estimated overall score from the estimated factor scores in the sentence-reading-aloud task.

and the RMSE are 0.83 and 0.39, respectively, which are close to the upper limit described in Sect. 3.2. We can also see that there is no remarkable difference among the three questions.

4. SENTENCE GENERATION TASK

4.1. Factors to be considered

In the sentence generation task, examinees speak the answer by a short sentence. The aim of the sentence generation task is to evaluate the ability to listen, speak, and generate sentences. It means that the overall score is affected by factors both in the quality of speech and in the content of speech.

Referring to the previous works [4] and [5] again, Listening (X_6), Representation (X_7), Grammar (X_8), and Vocabulary (X_9) were picked up as candidate factors, which belong to the content of speech, in addition to the five factors in the quality of speech described in Sect. 3.1. Listening (X_6) is the ability to understand a question correctly. Representation (X_7) is also the ability to generate a meaningful and rich sentence, for example, it may have a high score when using an expression fitting to a scene like honorific expressions.

4.2. Overall score estimation model

A subjective experiment was conducted to select the factors, which are effective for estimating the overall score, and to obtain the overall score estimation model for the sentence generation task. The experimental conditions are the same as those in Sect. 3.2.

Table 4. Correlation coefficient among all the factors in the sentence generation task.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1	1.00								
X_2	0.76	1.00							
X_3	0.76	0.88	1.00						
X_4	0.64	0.88	0.74	1.00					
X_5	0.55	0.49	0.43	0.50	1.00				
X_6	0.37	0.39	0.50	0.17	0.18	1.00			
X_7	0.62	0.68	0.69	0.51	0.38	0.75	1.00		
X_8	0.54	0.60	0.59	0.49	0.36	0.54	**0.86**	1.00	
X_9	0.63	0.60	0.62	0.42	0.39	0.69	**0.95**	0.84	1.00

Table 4 shows the correlation coefficients among all the factors. We can see that Representation (X_7) has a strong correlation with Grammar (X_8) and Vocabulary (X_9). The reason would be that the length of answer speech is too short to make a difference. It can also be seen that there is not a strong correlation between the factor in the quality of speech and the factor in the content of speech.

Using the same manner in Sect. 3.2, we selected Pronunciation (X_1) in the quality of speech, and Listening (X_6) and Representation (X_7) in the content of speech as effective factors, and defined the overall score estimation model in Fig. 1 by

$$\text{Overall score} = 0.43X_1 + 0.59X_6 + 0.40X_7 - 1.52. \quad (4)$$

We can see that the factors in the content of speech play an important role. It is valid that the selected factors correspond to the ability to speak, listen, and generate sentences.

Fig. 4 shows the relationship between the human overall score and the overall score estimated by substituting the human scores of each factor in Eq. (4). The human scores of each factor are the same as those used for obtaining Eq. (4). The correlation coefficient and the RMSE are 0.92 and 0.50, respectively. This corresponds to the upper limit of the estimation accuracy achieved by the proposed method.

4.3. Features to estimate the scores of each factor

For Pronunciation (X_1), we use the same features in the sentence-reading-aloud task, x_{1a} and x_{1b}, but they are calculated by using the time alignment to the output of dictation (large vocabulary continuous speech recognition) and the time alignment to the output of continuous phoneme recognition without any language limitation. Note that the dictation is performed to obtain the transcribed text of the answer speech, unlike the case in the sentence-reading-aloud task. Since any recognition system unavoidably produces recognition errors, we add the following feature.

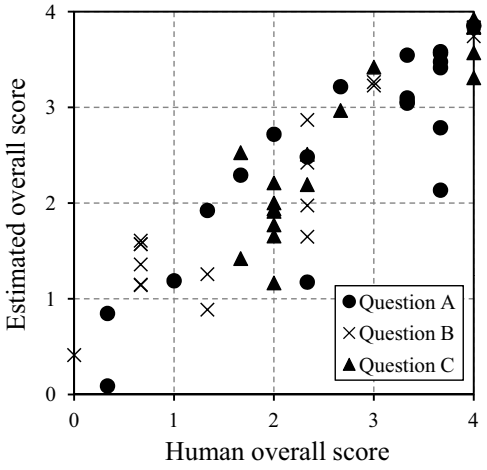

Fig. 4. Relationship between the human overall score and the estimated overall score from the human factor scores in the sentence generation task.

x_{1e} : Average of the word confidence scores, which are weighted by the word length.

The confidence score, which is outputted with the recognition result by the decoder, ranges from 0.0 (not reliable) to 1.0 (reliable)[11]. x_{1e} would become large when the recognition is done correctly.

We then discuss features for Listening (X_6) and Representation (X_7). In the sentence generation task, every question has its own keyword to be answered (hereafter referred to as 'important keyword'), for example, the keyword in a question about scheduling of a meeting is the date. Hence, the score of the factors is seriously affected by whether the answer speech includes the important keyword or not. Since the same tendency is observed for the predicate, we deal with the possible predicate as 'predicate keyword'. So we propose the following feature, which is calculated by using both a dictation-based keyword spotting method and a garbage model-based keyword spotting method.

x_{6a} (x_{7a}) : Ratio of the number of the keywords found by the two word spotting methods, to the total number of the keywords to be included in the answer speech. Note that the number of the important keywords was multiplied by 2.

The use of the two word spotting methods aims to avoid missing the keyword to be found. We also propose the following feature to catch the validity of the uttered sentence.

x_{6b} (x_{7b}) : Minimum edit distance between the result of dictation and every model answer text prepared in advance.

The edit distance represents the distance between two sentences and is defined as the minimum cost of transforming one sentence to another sentence. x_{6a} (x_{7a}) would become large for a meaningful sentence. Also, x_{6b} (x_{7b}) would become small for a valid sentence.

Table 5. Estimation accuracy for X_1, X_6 and X_7.

Factor	Correlation coefficient	RMSE
X_1	0.73	0.43
X_6	0.74	0.79
X_7	0.76	0.57

4.4. Effectiveness of the Proposed Method

We first examined the effectiveness of the features by estimating the scores of each factor. The speech samples are the same as those used for the subjective test in Sect. 4.2. To calculate the features, the Julius decoder [9] is used with the speaker independent triphone models, which were trained by using the CSJ native Japanese speech corpus [12] and then adapted to non-native speakers.

By the linear regression among the human scores of each factor and the values of the features, we defined the estimator for Pronunciation (X_1), Listening (X_6) and Representation (X_7) by

$$X_1 = 1.16x_{1c} - 3.11x_{1d} + 1.67x_{1e} + 1.37, \quad (5)$$

$$X_6 = 2.08x_{6a} - 0.72x_{6b} + 1.77, \text{and} \quad (6)$$

$$X_7 = 0.77x_{7a} - 1.53x_{7b} + 2.62, \quad (7)$$

respectively. We then estimated the scores of each factor to substitute the values of the features in Eqs. (5) to (7), respectively. The estimation accuracy is summarized in Table 5. We confirmed that a relatively accurate estimate can be obtained again.

Finally, we determined the overall score by substituting the scores of each factor estimated above in Eq. (4). Fig. 5 represents the relationship between the human overall score and the estimated overall score. The correlation coefficient and the RMSE are 0.82 and 0.72, respectively, which are relatively close to the upper limit describe in Sect. 4.2.

5. CONCLUSION

In this paper, we proposed the automatic scoring method considering the quality and the content of speech for the sentence-reading-aloud task and the sentence generation task in SCAT. We first analyzed the factors which contribute to the overall score. The effective factors were Pronunciation (X_1) and Fluency (X_4) for the sentence-reading-aloud task, and Pronunciation (X_1), Listening (X_6) and Representation (X_7) for the sentence generation task. We then described the ways to estimate the overall score and the factor scores and to extract

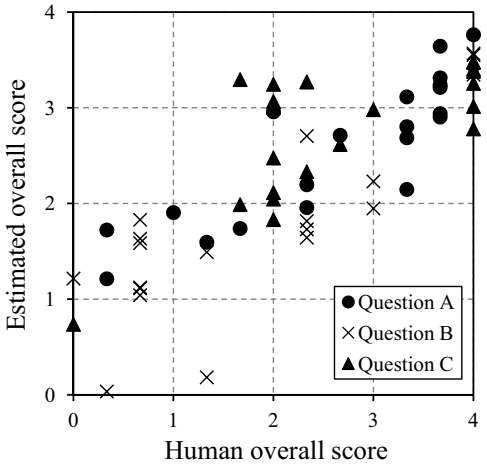

Fig. 5. Relationship between the human overall score and the estimated overall score from the estimated factor scores in the sentence generation task.

the features. Finally we verified the effectiveness of the proposed method. The correlation coefficient between the human overall score and the estimated overall score was 0.83 for the sentence-reading-aloud task and 0.82 for the sentence generation task. As future work, we plan to conduct an additional subjective experiment and then improve the performance and reliability of our method, comparing with conventional methods.

6. ACKNOWLEDGMENT

The authors would like to thank all members of the J-CAT project. Part of this research has been supported by KAKENHI (22242014).

7. REFERENCES

[1] J-CAT, http://www.j-cat.org/.

[2] S. Imai, S. Ito, Y. Nakamura, K. Kikuchi, Y. Akagi, H. Nakasono, A. Honda and T. Hiramura, "Features of J-CAT (Japanese computerized adaptive test)," Proc. 2009 GMAC Conference on Computerized Adaptive Testing, pp. 1–8, 2009.

[3] F. M. Lord, "Applications of Item Response Theory to Practical Testing Problems," Routledge, 1980.

[4] N. Fujishiro and I. Miyaji, "Effectiveness of Blended Instruction in Class on the Skills of Oral Reading and Speaking in English," Educational technology research, vol. 32(1·2), pp. 79–90, 2009.

[5] Versant English test, http://www.versanttest.co.uk/pdf/ValidationReport.pdf

[6] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech Communication, vol. 30, pp.95–108. 2000.

[7] M. Suzuki, Y. Qiao, N. Minematsu and K. Hirose, "Integration of multilayer regression with structure-based pronunciation assessment," Proc. INTERSPEECH2010, pp. 586–589, 2010.

[8] J. Doremalen, C. Cucchiarini and H. Strik, "Using Non-Native Error Patterns to Improve Pronunciation Verification," Proc. INTERSPEECH2010, pp. 590–593, 2010.

[9] A. Lee, T. Kawahara and S. Doshita, "An efficient two pass search algorithm using word trellis index," Proc. ICSLP1998, pp. 1831–1834, 1998.

[10] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, S. Sagayama, K. Itou, A. Ito, M. Yamamoto, A. Yamada, T. Utsuro and K. Shikano, "Free software toolkit for Japanese large vocabulary continuous speech recognition," Proc. ICSLP2000, pp. 476–479, 2000.

[11] A. Lee, K. Shikano and T. Kawahara, "Real-time word confidence scoring using local posterior probabilities on tree trellis search," Proc. ICASSP2004, pp. 793–796, 2004.

[12] T. Kawahara, H. Nanjo, T. Shinozaki and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," Proc. SSPR2003, pp. 135–138, 2003.