Bridging the Gap between L2 Research and Classroom Practice (2): Evaluation of Automatic Scoring System for L2 Speech

Yusuke Kondo[1], *Eiichiro Tsutsui*[2], *and Michiko Nakano*[3]

[1]Language Education Center, Ritsumeikan University, Japan
[2]International Center, Hiroshima International University, Japan
[3]Faculty of Education and Integrated Arts and Science, Waseda University, Japan
ykondo@fc.ritsumei.ac.jp, tsutsui@ic.hirokoku-u.ac.jp, nakanom@waseda.jp

Abstract

This paper introduces the construction, the implementation, and the evaluation of an automated scoring system for read-aloud speech of L2 learners'. In this system, evaluation scores given by trained human raters are predicted, based on the speech characteristics of learners in read-aloud speech.

Index Terms: automated scoring system, L2 speech, CEFR

1. Introduction

To assess speaking is one of the important objectives in second language (L2) assessment. In traditional speaking tests, learners' performance is assessed manually by trained raters on the respective criteria of proficiency standards. The process of this sort of test takes time. Therefore, several attempts have been made to construct automatic L2 speech scoring systems, and some of them are now in-service (e.g. [1] and [2]). In these studies, to automatically score the speech firstly learners' speech data are evaluated by raters and the speech characteristics (e.g. speech rate, number of silent pause, and quality of vowels) are measured, and secondly, the relationship between them is examined to obtain a prediction formula. Based on the formula, a new examinee's score is predicted by his/her speech characteristics.

This paper introduces the construction, the implementation, and the evaluation of an automated scoring system for read-aloud speech of L2 learners'. In this system, evaluation scores given by human raters are predicted, based on the speech characteristics of learners in read-aloud speech. The text that examinees were asked to read aloud is a fable from Aesop, "The North Wind and the Sun," and the evaluations given to examinees are categorical scores: A, B, and C.

2. Data

2.1. Speech data

Each informant out of 101 Asian English learners was recorded as they read a passage aloud. The group was composed of forty Japanese, seventeen Chinese, nineteen Korean, six Filipino, ten Thai, four Vietnamese, four Cambodians, and one Indonesian. These participants were either undergraduate or graduate students. Table 1 shows the key information of the participants.

Table 1 Key information of the informants

	M	*SD*	Range
Age	23.46	4.42	20
Study of English	11.88	5.41	29

All the recording was made in soundproof rooms in the universities which the participants belonged to. The informants were called in the room and given the instruction of recording individually. Their self-introductions without preparation were digital-tape recorded by using Roland R-09 and a condenser microphone, SONY ECM-MS957. In the recording, firstly the participants gave their self-introduction to an interviewer, and secondly read aloud the text. After the recording, the participants were given a small gift for their participation. It took about ten minutes for each participant to complete the recording.

The text that the informants read was a fable from Aesop, "The North Wind and the Sun," which is famous enough so that students at university level should know it. This passage was also used in the National Institute of Education Singapore corpus [3], and is used in the phonetic description of the International Phonetic Association.

This passage consists of 113 words: five sentences with a Flesch Reading Ease score of 79.9 and a Flesch-Kincaid Grade Level of 6.7. It contains almost all the vowels and consonants in English except for /ʒ/, /aʊ/, and /ɔɪ/ (the phonetic description is based on [4]).

2.2. Scoring by human raters

Five raters joined this evaluation; they were Japanese language teachers who had participated in the rater training, and their reliability had been examined in the evaluation of spontaneous speech [5]. The rater training was conducted according to Common European Framework of Reference (CEFR). In the rater training, the raters watched the video which depicted learners divided into six levels of CEFR, and discussed the characteristics of learners of each level. The raters evaluated all the speeches that were read by the 101 Asian English learners.

Evaluation items were selected from those in [6], and each item was thoroughly reviewed in order to make the items suitable in the evaluation of read-aloud speech. The items are depicted in Table 2.

Table 2 Evaluation items

1. Loudness	8. Sentence stress
2. Sound pitch	9. Rhythm
3. Quality of vowels	10. Intonation
4. Quality of consonants	11. Speech rate
5. Epenthesis	12. Fluency
6. Elision	13. Place of pause
7. Word stress	14. Frequency of pause

The evaluation scores were analyzed, based on Multifaceted Rasch Analysis (MFRA). Unreliable raters and items were detected based on their measures of infit, which were produced by MFRA. The infit measure "provides the size of the residuals, the differences between predicted and observed scores" [7]. The acceptable range of infit is "the mean ± twice the standard deviation of the mean score statistics" in cases when the population exceeds thirty. The analyses were repeated to meet this standard. In the present analysis, neither raters nor items exceeded the acceptable range.

2.3. Examination of the relationship between the evaluation scores and the speech characteristics

Five pilot studies were conducted to examine the relationships between the evaluation scores and speech timing control characteristics, pause control, vowel discrimination, reduced vowels, loudness, pitch, and pronunciation errors in read speech [8], [9], and [10]. Based on the results of the pilot studies, the relationship between speech characteristics and evaluation scores in the 101 read-aloud speeches was examined using multiple regression analysis (stepwise method).

The abilities estimated by MRFA are used as the criterion variable. The predictor variables are the two features that were adopted as indicators of evaluation scores in the analysis: the pruned syllables per second and the ratio of weak syllables to strong syllables. Pruned syllables per second was adopted as the index of speech rate. Pruned syllables per second are operationalized as follows:

$$S = (T – E) / TD \qquad (1)$$

where S is the speech rate index, T is the total number of syllables a learner uttered, E is the total number of unnecessary syllables (e.g., repetitions, fillers, and false starts), and TD is the total time duration [11]. The ratio of unaccented syllables to accented syllables is operationalized as follows:

$$R = A / U \qquad (2)$$

where R is the index of rhythm (namely the ratio of unstressed to stressed syllables), A is the average time duration of accented syllables, and U is the average time duration of unaccented syllables. This index is adopted from [12]. The average ratios of native English speakers are close to .5 or .4 [12].

The significance of the model was verified ($F_{(2, 98)} = 44.57$, $p < .01$, adjusted $R^2 = .47$). The correlation between the observed values and the predicted values is .69. Figure 1 is the scatter graph of the observed and predicted value, where the y-axis is the observed value and the x-axis is the predicted value.

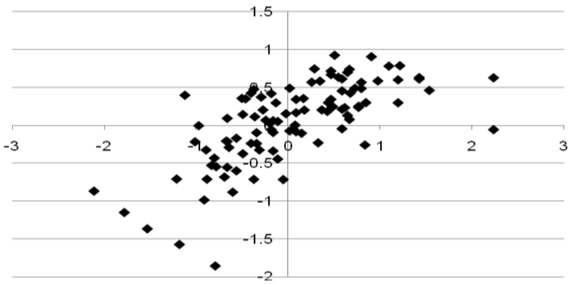


Figure 1 The Observed and Predicted Score

In this analysis, a high multiple correlation coefficient (.69) was obtained, though some outliers were found in the data. The goal of this study is to build an automatic speech evaluation system for L2 English learners. To obtain an accurate model it is possible to displace these outliers from our data by establishing a certain standard. However, from an educational point of view, we need to investigate objective measures to predict the evaluation scores of the outliers. Considering the coefficient of determination, however, we conclude that by using the learners' speech characteristics obtained in the previous analyses, we are able to replicate reliable and valid evaluation scores in the automatic L2 speech evaluation system.

3. The system

3.1. The structure

The automated scoring system to be implemented is a web-based system written the following procedures. Examinees read "The North Wind and the Sun" aloud on their client computers. Then, the recorded speech data are transferred to a server computer where the data are analyzed. Finally, the examinees receive feedback from the server computer on their client computer. Figure 2 depicts the automated scoring procedure.

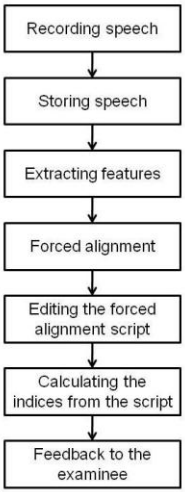
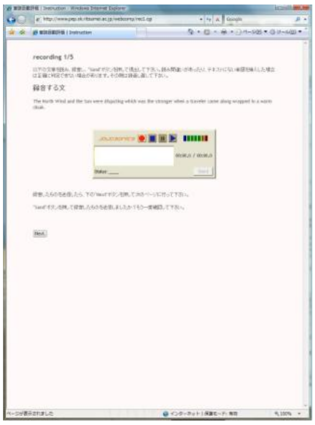


Figure 2 Procedure of Automated Scoring

The system records an examinee's speech using the Java applet, JavaSonics ListenUp [13], and this recorded speech is transferred to the sever computer and stored. Then, the speech is converted to the Hidden Markov Model Toolkit (HTK) format and analyzed. The results of forced alignment are edited to calculate the two indices: pruned syllables per second and the average ratio of weak syllables to strong syllables. Then, based on these two indices, the examinee's score is calculated and the feedback is sent to the examinee's computer. All of the processes are controlled by Perl scripts, including the JavaSonics ListenUp and HTK processes. The processes on the examinee's side are implemented with a web browser.

3.2. Test-taking procedure

To take the test through this system, firstly, examinees access the evaluation website, enter their names, and answer a questionnaire. They submit their answers and go to the instruction page. Secondly, on the instruction page, the examinees receive instructions on how to take the test, and they practice to record their speech. The whole passage that is to be read and its Japanese translation are provided on this page. After practice, they proceed to the recording page. In this test, they read "The North Wind and the Sun" aloud and record and submit their speech sentence by sentence. They record and submit their speech five times in total. Figure 3 shows the screenshot of the recording page.



recording 1/5

[illegible]

録音する文

The North Wind and the Sun were disputing which was the stronger when a traveler came along wrapped in a warm cloak.

[illegible]

Figure 3. The Screenshot of the Recording Page

4. Evaluation of the scoring methods

4.1. Level estimation based on NTT

The evaluation scores of the 101 read-aloud speeches were analyzed, based on Neural Test Theory (NTT) to estimate the examinees' levels. The proposed automated speech scoring system is a system that is meant to predict the evaluations given by human raters. Considering the reliability of human rating and the accuracy of its prediction by the system, it is reasonably appropriate to group examinees into three levels that correspond to the criterion given by CEFR: basic users, independent users, and proficient users. In this analysis, the levels are set up to three, and the fit of the data to the model is examined. The examinees were divided into three groups: thirty-six proficient users, thirty-one independent users, and forty-four basic users. The test fit indices indicate the data's goodness-of-fit to the model in NTT (χ^2_{156}=237.65).

4.2. The experiment

New speech data were obtained from twenty one Japanese university students. Their speeches were evaluated by three human raters and the proposed automatic evaluation system. The raters evaluated the twenty one learners' speeches according to CEFR, and gave ordinal evaluations: A, B, and C. To compute the ordinal evaluations by the system, three methods were used: Nearest neighbor (NN) method, k-NN method [14] and multiple regression. The reliability of these three scoring methods was examined in terms of the degree of the agreement with the evaluations by the human raters.

NN method and k-NN method are a pattern-recognizing technique used in image and speech recognition. In these methods, existing data are manually categorized based on their amount of characteristics beforehand, and a new data is grouped into the category according to its amount of the characteristics. In NN method, prototypes are decided by calculating the averages of amount of characteristics in each category of existing data, and a new data is grouped into a category that has the nearest prototype to the new data. In the present case, the levels of the speech data in Asian English speech database were decided based on the estimation by NTT, and the averages of the indices of speech rate and rhythm are calculated in each level. The averages are used as prototypes in each level. In scoring a new examinee, the two speech characteristics (the indices of speech rate and rhythm examined in 2.3) are measured, and the distance from the new examinee to the prototypes of three levels are calculated. The new examinee is grouped into the level that has the nearest prototype to the average of the new examinee.

In k-NN method, a new data is grouped into a category that has many data elements near to the new one. k is decided by an analyzer. If k is set to five, five data elements nearest to the new one are extracted, and the new data is grouped into a predominant category among the five data elements. In the present case, the levels of the speech data in Asian English speech database were decided by NTT. The two speech characteristics of a new examinee are measured, and five nearest data elements to the new data are selected in the existing data. A predominant level among the five data elements is assigned to the new examinee. For example, if the levels of five data elements nearest to a new data are A, A, B, C, and A, the new one is grouped into the level, A. Both in NN method and k-NN method, Euclidean distance is used as the distance metric.

In multi-regression, based on the abilities estimated by MFRA, the speech data were divided into three levels: twenty seven per cent of upper, forty six per cent of middle, and twenty seven per cent of lower levels, and the high and low limits of the scores in each level were calculated. The two speech characteristics, the indices of speech rate and rhythm, of a new examinees are measured, and the new examinee's score is predicted adopting the multiple regression formula obtained in the correlation study. The examinee is grouped into a level whose range includes the examinee's score.

To the degree of agreement of the three scoring methods with the human raters, two methods were adopted: Fleiss' kappa and the correlation coefficients among the human raters. The degrees of agreement were examined based on Fleiss' kappa among the scores given by the human raters and the three sorts of scores computed by the automated scoring system.

Fleiss' kappa [15] is a measure of inter-rater reliability for assessing the degree of agreement when more than three raters evaluate performance with a fixed number of categories [16]. The interpretation of this index is somewhat controversial, because it depends on the number of raters, categories, and examinees. The Fleiss' [15] interpretation of kappa is as follows: kappa below .40 represents "poor agreement beyond chance, the value above .75 represents "excellent agreement beyond chance", and the value between .75 and .40 represents "fair to good agreement beyond chance". Table 3 shows the Fleiss' kappa among the human raters and the three sorts of the scoring methods. Each value is the kappa among one scoring method and the three human raters. The highest value was obtained by NN method. Although all kappa fall into the range of "fair to good agreement beyond chance" according to the Fleiss' interpretation, NN method obtained the highest kappa. Table 4 shows the correlation coefficients among the three human raters and the three scoring methods. NN method obtained the highest correlation coefficients with all the raters.

Table 3 Fleiss' kappa among the raters and the methods

Method	κ
NN method	.58
k-NN method	.42
Multiple regression	.49

Table 4 Correlation coefficients among the raters and the methods

	NN method	k-NN method	Multiple regression
Rater 1	.81	.52	.67
Rater 2	.69	.61	.61
Rater 3	.58	.52	.54

Table 5 shows the correlation coefficients between the human raters and the system (NN method). The correlations among the human raters were fairly high, and compared to the correlation among the human raters, relatively low correlation coefficients were found between the human raters and the system. Nevertheless, substantial correlation coefficients among the human raters and the system were found in this study. To obtain the average of the correlation coefficients above, z-transformed values were computed. The average of the inter-rater reliability in this evaluation is .79.

Table 5 Correlation coefficients between the raters and the system (NN method)

	System	Rater 1	Rater 2	Rater 3
System	1	.81	.69	.58
Rater 1		1	.83	.80
Rater 2			1	.89
Rater 3				1

5. Discussion and conclusions

This paper introduced the automatic L2 speech evaluation system that predicted the evaluation by human raters by using learners' speech characteristics from the read-aloud speech of English learners. In this system, an examinee is categorized into one of the three levels based on the speech data of 101 Asian English learners that were given evaluation scores by trained human raters. The evaluation in the system is determined by the two predictor variables: pruned syllables per second and the ratio of weak syllables to strong syllables. The ability of these variables to predict evaluation scores was verified. The system operates via the internet, and an examinee may take the test in any place that the internet is available.

The evaluation scores produced by the system were examined. The degrees of agreement by the Fleiss' kappa showed that though the degree of agreement was the highest among only the human raters (.75), the degrees of agreement of the system with the human raters were sufficiently high (.70, .60, and .60). Furthermore, the average of the correlation coefficients among the human raters and the system was fairly high (.79). Judging from the results of the experiments, it appears to be possible that we may obtain reliable evaluation scores by using the automatic L2 speech evaluation system. The system was constructed for experimental use and is not adequate for simultaneous access, but if the part of the system is improved, it can be adapted for practical use. The practical application of this system can be an effective tool to assess second language learners' performance. The results of this experiment indicate the possibility that the evaluation of read-aloud-speech performed by trained human raters can be predicted by learners' speech characteristics which computers are capable of calculating. In other words, we can obtain reliable evaluation scores in read speech by using computers.

Fleiss' kappa was adopted as the index of rater agreement. Although the agreement was the highest among the raters, substantial agreement was obtained between the human raters and the system. Perfect agreement is difficult to achieve in the performance assessment, as was indicated by the Fleiss' kappa among the three human raters (.75). Furthermore, the average of the correlation coefficients among the human raters and the system is .79, which indicates the high reliability of this evaluation. The evaluation given by the human raters in this experiment was an overall evaluation of read-aloud speech, and the evaluation scale was a 3-point scale (A, B, and C). Hence, we cannot make a simple comparison between the results of the present study and those of previous studies, but the average of the correlation among the human raters and the system falls into an acceptable range of inter-rater reliability.

6. Acknowledgements

This work was supported by KAKENHI (Grant-in-Aid for Young Scientists (B) 21720209).

7. References

[1] Bernstein, J., De Jong, J. Pisoni, D., & Twonshend, B. (2000). Two experiments on automatic scoring of spoken language proficiency. *Proceedings of InSTIL 2000*, 57-81.

[2] Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). Automated scoring of spontaneous speech using SpeechRater v1.0. (ETS Research Report No. RR-08-62). Princeton, NJ: ETS.

[3] Deterding, D., & Ling, L. E. The NIE corpus of spoken Singapore English. Deterding, In D., Brown, A., & Ling, L. E. (Eds). (2005). *English in Singapore*. Singapore: McGraw-Hill Education.

[4] Jones, D. (2003). *Cambridge English pronouncing dictionary*. Cambridge: CUP.

[5] Nakano, M., Kondo, Y., Tsubaki, H., & Sagisaka, Y. (2008). Rater Training Effect in L2 and EFL Speech Evaluation. The 8th Phonetics Conference of China and the International Symposium on Phonetic Frontiers.

[6] Yashiro, K., Araki, A., Higuchi, Y., Yamamoto, S., & Komissarov, K. (2001). *Ibunka communication workbook*. [A workbook for cross-cultural communication]. Tokyo: Sanshusha.

[7] McNamara, T. F. (1996). *Measuring second language performance*. Essex: Pearson Education Limited.

[8] Kondo, Y., & Nakano, M. (2009). Construction and implementation of automatic L2 speech evaluation system. *Proceedings of 14th Conference of Pan-pacific Association of Applied Linguistics*, 33-38.

[9] Kondo, Y., Tsutsui, E., Nakano, M., Tsubaki, H., Nakamura, S., & Sagisaka, M. (2007). "The relationship between subjective evaluation and objective measurements in Second language oral reading" [Eigo gakushusha ni yoru ondoku ni okeru shukanteki hyoka to kyakkanteki sokuteichi no kankei]. *Proceedings of the 21st General Meeting of the Phonetic Society of Japan*. 51-55.

[10] Kondo, Y., Tsutsui, E., Tsubaki, H., Nakamura, S., Sagisaka, Y., & Nakano, M. (2007). Examining predictors of second language speech evaluation. *Proceedings of 12th Conference of Pan-Pacific Association of Applied Linguistics*, 176-179.

[11] Riggenbach, H. (1991). Toward an understanding of fluency: A micro-analysis of nonnative conversations. *Discourse Processes*, 14. 423-441.

[12] Derwing, T. M., Rossiter, M. J., Munro, J. M., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*. 54, 4. 655-679.

[13] Mobiller, Inc. (2006). ListenUp SDK [Computer program]. http://www.javasonics.com/

[14] Shakhnarovich, G., Darrell, T., & Indyk, P. (eds.). (2006). *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*. Cambridge, MA: The MIT Press.

[15] Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. Hoboken: Wiley-Interscience.

[16] Gwet, K. (2001). *Handbook of Inter-Rater Reliability*. StatAxis Publishing Company.