A method for reducing burden imposed on human raters in the construction of automated scoring systems for second language learners' speech

Yusuke Kondo

Open Education Center, Waseda University

yusukekondo@aoni.waseda.jp

Abstract

Attempts have been made to construct automated scoring systems for second language learners' speech. In the initial stage of the construction of these systems, the relationship is investigated between the scores by human raters and speech characteristics that are measurable by computer in order to obtain prediction formulae: Once we obtain the formulae, examinees' scores can be predicted using speech characteristics. Although the computerized assessment is proposed as one of the solutions to reduce the raters' burden, the initial stage of the system construction requires a large amount of learners' speech data with the scores given by human raters. The raters need to evaluate a large set of speech samples. To solve this problem, this study proposes a method for predicting the scores of a large set of unscored speech data by a small set of speech data with the human rating. The speech data used in this study are 101 read-aloud speeches given by Asian learners of English. Using two speech characteristics of five speeches randomly selected, the scores of the remaining 86 speeches are predicted, based on Expectation-Maximum (EM) algorithm. The moderate correlation was found between the scores given by the human raters and the ones predicted by the algorithm (around .60).

Keywords

L2 speech assessment, automatic scoring, EM algorithm

1 Introduction

In language tests to assess L2 speaking skills in general, examinees are asked to introduce themselves, to describe some pictures, or to discuss general issues, and raters evaluate the examinees' speech, as in ACTFL OPI, STEP TESTS, and various versions of Cambridge Proficiency Tests, etc. In these tests, their oral performance is assessed manually by trained raters based on the respective criteria of proficiency standards. Before conducting this kind of test, we usually conduct workshop several times for our raters to arrive at good inter-rater agreement. The evaluation scores are analyzed to exclude unreliable raters on the basis of some statistical models. The use of rater will introduce another problem to the assessment: Inconsistency of assessment and investment of time. It is often hoped that as a solution, automated scoring of L2 speech be built to predict the evaluations by human raters.

One of the purposes to construct an automated scoring system for L2 speech is to reduce burden imposed on human raters. However, we could not reduce burden imposed on human raters in the process of the construction of the system. Once we have constructed the automated scoring system, we will achieve consistent assessment continuously by using the developed system. However, we burden raters with a number of scoring in the process of constructing the system. Therefore, it is hoped that a method be formulated for reducing the burden imposed on human raters in the process of the construction of the automated scoring system. In this paper, a method to predict the scores of unscored speeches applying Expected-Maximization (EM) algorithm.

Section 2 provides a brief summary of the existing automated scoring system for L2 read-aloud speech developed in Kondo (2010), pointing out the problem related to the process of human rating. Section 3 describes the proposed method and reports the results. Lastly, in Section 4, the usefulness of the proposed method will be discussed.

2 The automated scoring system for L2 read-aloud system in Kondo (2010)

This section introduces the existing automated scoring system for L2 read-aloud speech developed in Kondo (2010) and presents a problem in construction of the automated scoring system.

2.1 Examination of the relationship between speech characteristics and human rating

The system by Kondo (2010) predicts an examinee's score on the basis of the relationship between the human scoring and the speech characteristics in the existing data. The relationship was examined between the evaluation scores and the speech characteristics realized in read-aloud speech (Kondo, Tsutusi, Nakano, Tsubaki, Nakamura & Sagisaka, 2007; Kondo, Tsutsui, Tsubaki, Nakamura, Sagisaka & Nakano, 2007). The examined speech characteristics were number of non-lexical pause, number of silent pause, duration of non-lexical pause, duration of silent pause, mean length of run, number of syllable unneeded, pruned syllable per second, and the ratio of weak syllable to strong syllable. These objective measures were selected from Munro and Thomson (2004); Trofimovich and Baker (2006 and 2007); Riggenbach (1991); and Towell, Hawkins, and Bazergui (1996). Through the pilot studies of correlation (Kitagawa, Kondo & Nakano, 2007; Nakano, Kondo & Tsutsui, 2008), we found the two independent predictors of the evaluation scores: the pruned syllable per second and the ratio of weak syllable to strong syllable.

Pruned syllables per second are operationalized as follows:

$$S = (T - E) / D \tag{1}$$

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 D is the total time duration (Riggenbach, 1991). The ratio of unaccented syllables to accented syllables is operationalized as follows:

$$R = A / U \tag{2}$$

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 Derwing, Rossiter, Munro, and Thomson (2004). The average ratios of native English speakers are close to .5 or .4 (Derwing, Rossiter, Munro, and Thomson, 2004).

The relationship between human scoring and the two speech characteristics was examined by multi-regression analysis (stepwise method): the criterion variable is the evaluation score; and the predictor variables, the pruned syllable per second and the ratio of weak syllable to strong syllable. 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.

2.2 Scoring method

The evaluation scores were standardized by Latent Rank Theory (LRT: Shojima, 2008) to estimate the examinees' ranks. LRT is a test theory which adopts the mechanism of self-organizing map (SOM: Kohonen, 2000). In LRT, ordinal scale is assumed as latent scale. Examinees are grouped into some ranks a test developer sets up according to the probability estimated. Because LRT estimates the probability which levels examinees are grouped into based on raw scores, the correlation is fairly high between raw scores and the levels estimated in LRT. The computational procedure of LRT is identical to that of SOM.

In LRT, the response data of a new examinee are examined and are categorized into a latent rank that has the closest vector. Then, based on the categorized response data of the new examinee, all the reference vectors are updated. The outline of computational procedure of LRT is outlined below (Shojima, 2008):

```
For ($t = 1; t \leq T; t = t + 1$)                                                  (L1)
Obtain $U^{(t)}$ by randomly sorting the row vectors of U.                          (L2)
For ($h = 1; h \leq N; h = h + 1$)                                                  (L3)
      Input $u_h^{(t)}$, the h-th row vector of $U^{(t)}$, and select the rank with     (L4)
      the closest reference vector in terms of the discrepancy function $d$.
      Obtain $V_h^{(t)}$ after updating the reference vectors of the winner and      (L5)
      neighboring nodes.
   $V(t+1) \Leftarrow V_n^{(t)}$                                                     (L6)
```

where T is learning time set by an analyzer; N, sample size; U, the response data of examinees, U = {u_i}(i = 1,…, *N*); and V, reference vector with the number of item × the number of the latent lank. The procedure of (L1) requires that of from (L2) to (L6) repeatedly until t equals T. Similarly the procedure of (L3) requires that of (L4) and (L5) repeatedly until h equals *N*. The discrepancy function *d* in (L4) is determined by the following formula (Shojima, 2007: 3):

$$\mathrm{R_w} : \mathrm{w} = \arg\min\nolimits_{\mathrm{q} \in \mathrm{Q}} \parallel \mathrm{v}_\mathrm{q}^{(\mathrm{t})} - \mathrm{u}_\mathrm{h}^{(\mathrm{t})} \parallel^2 \tag{3}$$

where Q is the number of latent ranks set by the analyzer, and $\mathrm{v}_\mathrm{q}^{(\mathrm{t})}$ is the reference vector of a rank at the t-th period, $\mathrm{u}_\mathrm{h}^{(\mathrm{t})}$ is the response data of an examinee.

The speech data in the present study were ranked into three, A, B and C. In Figure 1, the ranked speech data of the learners are identified with the values of pruned syllable per second and the average ratio of weak syllable to the strong syllable. The x-axis indicates the value of the average ratio of weak syllable to the strong syllable; and the y-axis, the values of pruned syllable per second. The values of the average ratio of weak syllable to the strong syllable are inverted (the plotted values are 1 minus the original values) for the clear picture. Although some outliers and a multi-occupied area by all the three ranks are found, the areas of each rank can be specified to some extent. The averages of the two values were calculated in each category, and plotted. The bigger indicators are the averages (prototypes) of the two values in each category.

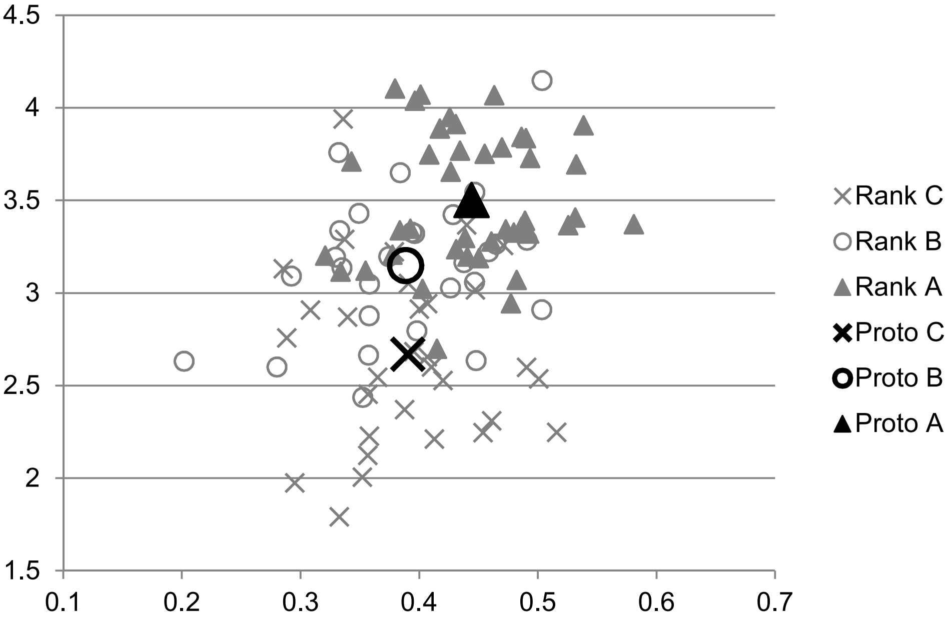

Figure 1: Scatter Graph for the Values of Pruned Syllable per Second and the Average Ratio of Weak Syllable to the Strong Syllable in Each Category

The system adopted Nearest Neighbor method (Shakhnarovich, Darrell, and Indyk, 2006) to decide the rank of a new examinee. A new examinee's category is determined on the basis of the Euclidean distances to the prototypes in each category. The distances to each prototype are calculated in the equation below:

$$\mathrm{D(x,p)} = \sqrt{(\mathrm{p_1} - \mathrm{x})^2 + (\mathrm{p_2} - \mathrm{y})^2} \tag{4}$$

where p_1 is the average of pruned syllable per second in a category, and x, the pruned syllable per second of a new examinee's; and p_2 is the average of the ratio of the weak syllable to the strong syllable, and y, the ratio of the weak syllable to the strong syllable of a new examinee's. Comparing the three distances of the new examinee's values to each prototype: Rank A, B and C, the examinee is given the category whose prototypes obtain the nearest distance to the new data.

2.3 Evaluation of the system

An experiment was conducted to evaluate the accuracy of the scores produced by the system. Participants of this experiment were twenty Japanese learners of English and three raters. The raters were the Japanese language teachers of English who received rater training according to Common European Framework of Reference (CEFR, Council of Europe, 2001). The raters evaluated the twenty learners' speeches according to the criterion, CEFR and gave the categorical evaluations: A, B and C. The degree of agreement was examined in the evaluation scores among the human raters and the automatic evaluation system based on Fleiss' kappa (Fleiss, 1971).

Fleiss' kappa is a measure of inter-rater reliability for assessing the degree of agreement when more than three raters evaluate performance using evaluation items with a fixed number of categories (Gwet, 2001). Table 1 shows the Fleiss' kappa of the evaluation scores. The indices were calculated four times. In each time, one of the raters was excluded. By comparing these indices, the rater lowering the degree of agreement can be detected. For example, the kappa in the second row indicates the rater agreement among Rater 1, 2 and the automatic evaluation system, excluding Rater 3. The kappa in the lowest row indicates the rater agreement among all the raters: Rater 1, 2, 3 and the system.

Table 1: Fleiss' kappa among the raters and the system

Raters	κ
Rater 1, 2, and the system	.70
Rater 1, 3 and the system	.60
Rater 2, 3, and the system	.60
Rater 1, 2, and 3	.75
ALL	.66

Table 2 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.

Table 2: The correlation coefficients between the human raters and the system

	The system	Rater 1	Rater 2	Rater 3
The system	1	.81	.69	.58
Rater 1		1	.83	.80
Rater 2			1	.89
Rater 3				1

Though the agreement was the highest when the system was excluded from the raters, we obtained substantial agreement between the human raters and the system.

2.4 Problem in the system

One of the purposes to construct an automated scoring system for L2 speech is to reduce burden imposed on human raters. However, we impose burden on human raters in the process of the construction of the existing system. In Kondo (2010), workshops were conducted three times for the raters to arrive at good inter-rater agreement and analyzed the evaluations scores to exclude unreliable raters on the basis of two statistical techniques, Multi-faceted Rasch Analysis (Linacre, 1994) and Generalizability (Theory Brennan, 1992). Then, the trained raters evaluated 101 speeches given by Asian learners of English. We burden raters with a number of scoring in the process of constructing the system. Only 101 speeches are needed to be scored in Kondo (ibid), but if we construct the system adopting other elicitation task, then we need human rating to investigate the relationship between the score and the speech characteristics in that task. Therefore, it is hoped that a method be formulated for reduce the burden imposed on human raters in the process of the construction of the automated scoring system.

3 Prediction by EM algorithm

The speeches used in the present study are read-aloud speeches given by 101 Asian learners of English, all of which were scored by trained human raters and were categorized into three levels: A, B and C. In this data set, the relationship has been examined between the score and the speech characteristics: the two speech characteristics, the indices of speech rate and rhythm, were found to be statistically significant predictors of the scores. Using the two speech characteristics of five speech samples randomly selected from each level, the scores of the remaining speech samples are predicted, based on Expectation-Maximum (EM) algorithm (Dempster, Laird, Rubin, 1977), which requires two default values: mean and standard deviation. In this procedure, posterior probabilities, which are probability that a speech is categorized into A, B and C, are given to each speech, and the level of a speech is decided by comparing three posterior probabilities: if a speech obtains probability of .23 for A, .66 for B and .11 for C as their respective posterior probabilities, then the speech is categorized into rank B for which probability is the highest among three ranks.

In the present study, two-dimensional normal distribution with two variables, the indices of speech rate and rhythm is assumed. Multiply the mixing rate (ξ) by likelihood of being extracted from each distribution. For every single data, this value is calculated in Formula 5 for each rank.

$$\mathrm{f}(x_i|\mu_k,\sigma_k,\rho) = \xi \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{\left[\frac{1}{1-\rho^2}\left\{\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}\right]} \quad (5)$$

The ratios of the value below are regarded as the probability of the belongingness (z) which is obtained in Formula 6.

$$z_{ik} = \frac{\mathrm{f}(x_i|\mu_k,\sigma_k,\rho)}{\sum_l \xi_l \mathrm{f}(x_i|\mu_l,\sigma_l,\rho)} \quad (6)$$

If we stop here, we decide the score based on the probability of the belongingness. If an examinee get the probabilities: 0.3 for A, 0.5 for B, and 0.2 for C, the examinee is categorized into B.

The average of the likelihood (Q) is used as the goodness of fit. If the likelihoods of x are A: 0.03, B: 2.49, and C: 0.54, and the probability of the belongingness of x are A; 0.03, B: 0.80 and C: 0.17, then, the average of the likelihood is obtained in 0.01×0.03 + 0.80×2.49 + 0.17×0.54 = 2.84. Q is obtained in Formula 7. This process is called E-step.

$$\mathrm{Q}(\mathrm{x}|\xi,\mu,\sigma,\rho,z) = \sum_i \sum_k z_{ik} log \xi_k f\left(x_i|\mu_{k,}\sigma_k,\rho_k\right) \quad (7)$$

Applying a newly obtained variable z, the average μ, the mixing rate ξ, the standard deviation σ and the correlation coefficient ρ are updated in the formula 8, 9, 10 and 11 respectively. This process is called M-step.

$$\mu_k = \frac{\sum_i z_{ik} x_i}{\sum_i z_{ik}} \quad (8)$$

$$\xi_k = \frac{1}{N} \sum_i z_{ik} \quad (9)$$

$$\sigma_k = \sqrt{\frac{\sum_i z_{ik}(x_i-\mu_k)^2}{\sum_i z_{ik}}} \quad (10)$$

$$\rho_k = \frac{\sum_i z_{ik}(x_{ij}-\mu_{kj})(x_{ij}-\mu_{kl})}{z_{ik}\sum_i z_{ik}(x_i-\mu_k)^2 \sum_i z_{ik}(x_i-\mu_k)^2} \quad (11)$$

In the present study, each five sample speeches are randomly selected from each category twice, and the averages, standard deviations, correlation between two variables, and mixing rates are calculated (Table 3 and 4).

The actual data was plotted in Figure 2, and the data predicted by EM algorithm was shown in Figure 3 and 4. The first prediction by EM algorithm converged at 10th time, and the second one, at fourth time. The correlation of the first prediction with the scored data by human raters is .64 (Spearman's ρ), and that of the second prediction with the scored data by human raters is .60 (Spearman's ρ).

Table 3: The initial values for the first prediction

	Rank A	Rank B	Rank C
Average of rhythm	0.43	0.41	0.35
Average of speech rate	3.47	3.18	3.02
SD of rhythm	0.07	0.08	0.05
SD of speech rate	0.23	0.19	0.57
Correlation	0.02	-0.33	-0.41
Mixing rate	0.33	0.33	0.33

Table 4: The initial values for the second prediction

	Rank A	Rank B	Rank C
Average of rhythm	0.45	0.40	0.36
Average of speech rate	3.50	3.06	2.78
SD of rhythm	0.05	0.04	0.06
SD of speech rate	0.32	0.41	0.53
Correlation	0.02	-0.63	-0.41
Mixing rate	0.33	0.33	0.33

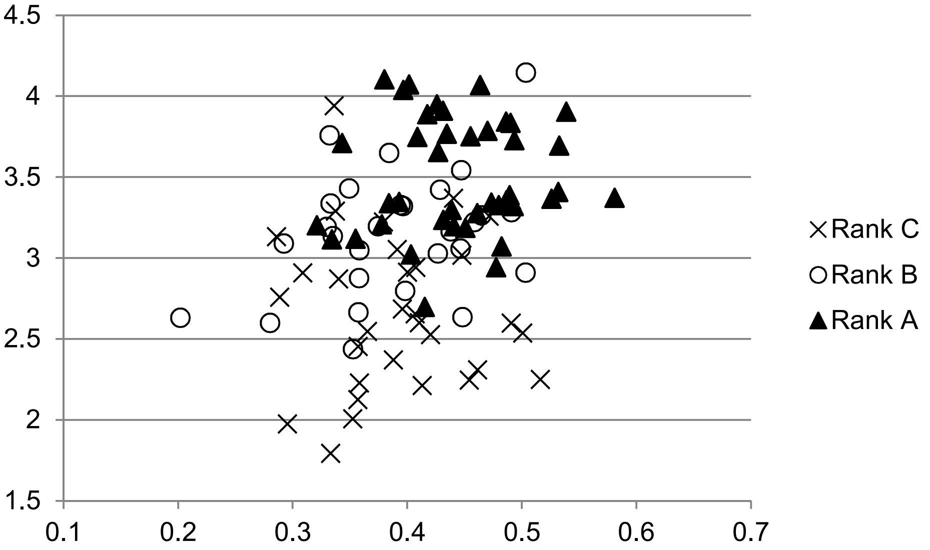

Figure 2: Scored data by human raters

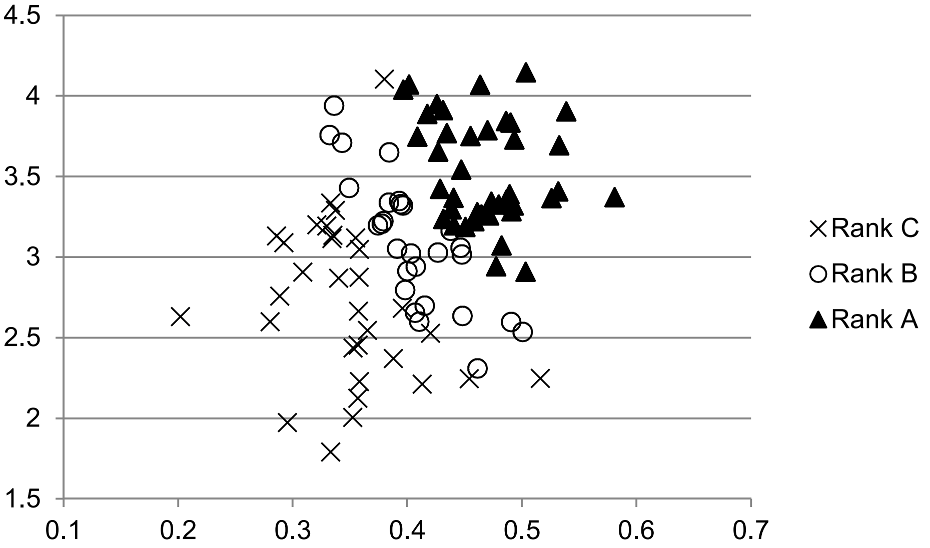

Figure 3: Prediction by EM Algorithm (1)

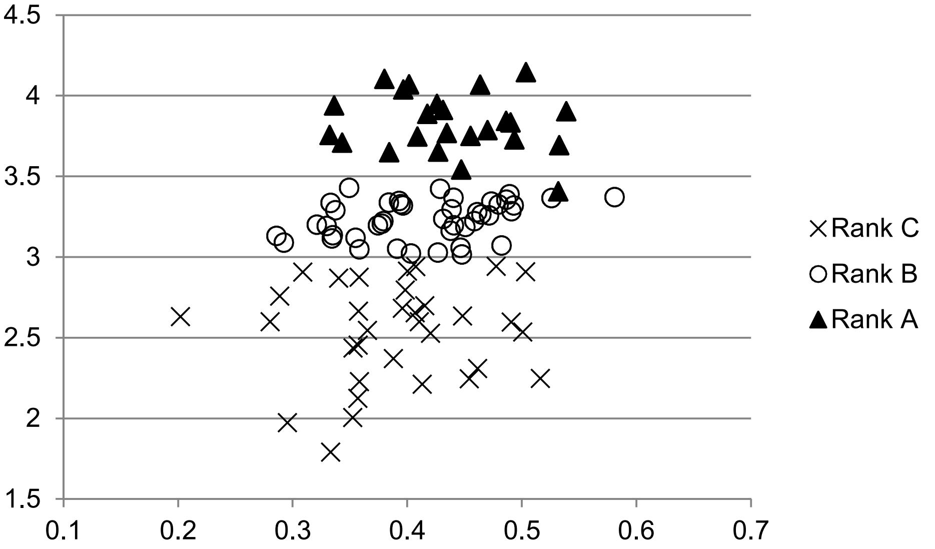

Figure 4: Prediction by EM Algorithm (2)

4 Discussion and conclusion

EM algorithm is one of the good methods to predict scores of unscored speech data by using a small set of the data. The indices adopted to examine the inter-rater reliability are different in the evaluation of the existing system and that of the present method (Fleiss' κ and Spearman's ρ). In the present study, moderate correlation was found between the first prediction by EM algorithm and the scored data by human raters (.64). As the results of the first and second predictions indicate, the initial values are very important in this method. It seems to be fairly difficult to obtain good results by using only two variables. Two or more variables are needed to obtain good results.

Acknowledgement
This work was supported by Grant-in-Aid for Scientific Research (C) (23520721).

References

Brennan, R. L. (1992). Generalizability Theory. ITEMS: The Instructional Topics in Educational Measurement Series. Module 14. Madison: NCME.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: CUP.

Dempster, A.P., Laird, N.M., Rubin, D.B., (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1): 1–38.

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.

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

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

Kitagawa, A., Kondo, Y., & Nakano, M. (2007). Does vowel quality matter? *Proceedings of 12th Conference of Pan-Pacific Association of Applied Linguistics*, 224-227.

Kohonen, T. (2000). *Self-organizing maps*. New York: Springer.

Kondo, Y. (2010). *The development of automatic speech evaluation system for learners of English*. Unpublished doctoral dissertation, Waseda University, Tokyo, Japan.

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.

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.

Linacre, J. M. (1994). *Many-Facet Rasch measurement*. Chicago: Institute for Objective Measurement, Inc.

Nakano, M., Kondo, N, & Tsutsui, E. (2008). Fundamental Research on Automatic Speech Evaluation. *9th APRU Distance Learning and the Internet Conference--New Directions for Inter-institutional*

Collaboration: Assessment & Evaluation in Cyber Learning. 207-212.

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

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

Shojima, K. (2008). Neural test theory: A latent rank theory for analyzing test data. *DNC Research Note*. 08-01.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistic*, 17, 1. 84-119.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, 28, 1-30.

________. (2007). Learning prosody and fluency characteristics of second language speech: The effect of experience on child learners' acquisition of five suprasegmentals. *Applied Psycholinguistics*, 28, 251-276.