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

This study proposes a method for predicting the scores of a large set of unscored speech data by a small set of speech data scored by human raters. Using two speech characteristics of ten speeches randomly selected from each level, the scores of the remaining 91 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). EM algorithm is one of the best methods to predict the scores of a large set of unscored speech data by using the information of a small set of scored speech data, but the results completely depends on the initial values in the present data set. We should examine methods for determining the initial values to obtain desired results.

Keywords

L2 speech assessment, automatic scoring system for L2 speech, EM algorithm

Introduction

One of the purposes to construct an automated scoring system for L2 speech is to reduce burden imposed on human raters. The implementation of speaking tests, such as an interview or a picture description task takes time. In order to reduce such cost, it is often hoped that as a solution, automatic L2 speech evaluation system be built to predict the evaluations by human raters. However, we could not reduce burden imposed on human raters in the process of the construction of the existing system. We conducted the workshop three times for our raters to arrive at good inter-rater agreement and analyzed the evaluations scores to exclude unreliable raters on the basis of two statistical techniques, MFRA and Generalizability Theory. Then, we asked the trained raters to evaluate 101 speeches given by Asian learners of English. 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. Only 101 speeches are needed to be scored in this study, 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.

1 The speech data and the existing system

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 ten speeches randomly selected from each level, the scores of the remaining speeches are predicted, based on Expectation-Maximum algorithm, 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 .23 for A, .66 for B and .11 for C as their respective posterior probabilities, then the speech is categorized into level B.

The existing system predicts an examinee's score by the nearest neighbor method. The averages of the two variables in each category: A, B and C are calculated, and a new examinee's category is determined based on the Euclidean distance: a

category that has the closest distance to the new examinee is assigned to the examinee.

Twenty new examinees are scored by this method and three human raters. The results of the inter-rater reliability are shown in Table 1.

Table1: Inter-rater reliability (Fleiss' κ)

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

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 Prediction by EM algorithm

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 2 and 3).

Table 2: The initial values for the first prediction

	Rank A	Rank B	Rank C
Average 1	0.43	0.41	0.35
Average 2	3.47	3.18	3.02
SD 1	0.07	0.08	0.05
SD 2	0.23	0.19	0.57
Correlation	0.02	-0.33	-0.41
Mixing rate	0.33	0.33	0.33

Note. 1 is the index of rhythm, and 2 is the index of speech rate

Table 3: The initial values for the first prediction

	Rank A	Rank B	Rank C
Average 1	0.45	0.40	0.36
Average 2	3.50	3.06	2.78
SD 1	0.05	0.04	0.06
SD 2	0.32	0.41	0.53
Correlation	0.02	-0.63	-0.41
Mixing rate	0.33	0.33	0.33

Note. 1 is the index of rhythm, and 2 is the index of speech rate

The actual data was plotted in Figure 1, and the data predicted by EM algorithm was shown in Figure 2 and 3.

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).

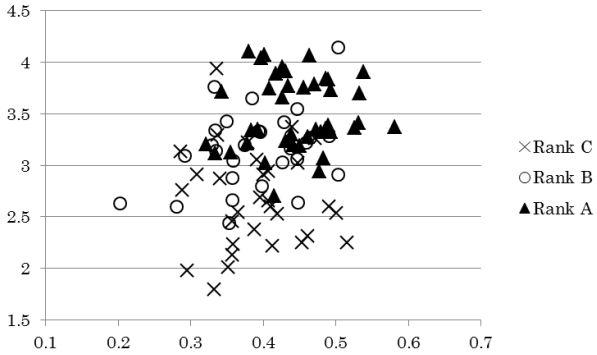

Figure 1: Scored data by human raters

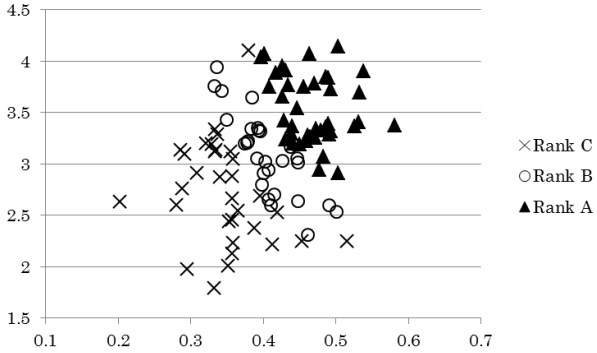

Figure 2: Prediction by EM Algorithm (1)

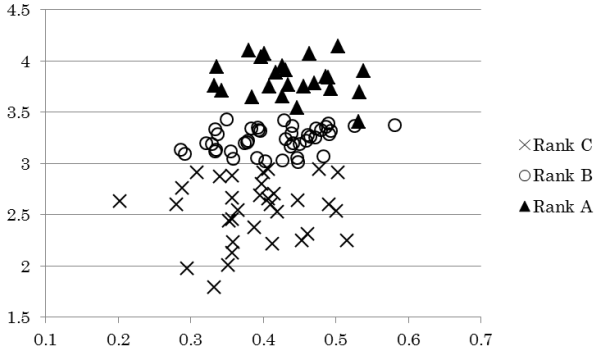

Figure 3: Prediction by EM Algorithm (2)

3 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.