Towards a new model of speech evaluation system within the framework of World Englishes in Asia

Yusuke Kondo[1], Norifumi Ueda[1], Eiichiro Tsutsui[2], Kazuharu Owada[3], Michiko Nakano[1]

[1]Waseda University, [2]Hiroshima International University,
[3]Ritsumeikan University

Abstract

In the field of speech recognition, attempts have been made to develop an automatic evaluation system for spoken language of second language (L2) learners. The models adopted in the previous studies are a multiple regression model where naturalness of speech is estimated by differences from the speech of the native speakers in the timing characteristics of overall sentence, individual content and function words, pauses, etc. The explained variable is the naturalness evaluation by a few native speakers of the target language. Although the estimation accuracy of the models was validated using open data, some need to be improved in selecting variables and the evaluation by raters. Furthermore, reflecting the situation where English is now used as an international language, it might be controversial that measuring the differences in speech between L2 learners and the native speakers. To revise the models, we collected the data from L2 learners of English with various first language backgrounds in Asia: Korean, Tagalog, Chinese, Malay, Thai, and Japanese. The learners read out a fairy tale, "The North Wind and the Sun", and were digital-tape recorded. In the current study, we adopt not only the timing characteristics analyzed in the previous studies, but also prosodic features. We investigate the relationship between the existing and new explanatory variables, and propose optimized predictor variables.

1. Introduction

This study is based on Muto, Sagisaka, Naito, Maeki, Kondo, and Shirai (2003) where they investigated the relationship between speech timing control characteristics and evaluation of Japanese learners of English in terms of rhythmic feature. Their participants read out several sets of sentences, and their raters evaluated the learners' speech in terms of rhythmic feature. Muto et al (ibid) investigated how well the objective measurements predict the raters' evaluation using multiple regression analysis.

The present study investigated the relationships between the objective measurements and

the raters' evaluation within the same framework as in Muto et al (ibid), but the present study differs from Muto et al (ibid) in five conditions. (1) The evaluation scores were calculated based on multi-faceted Rasch model, while the previous study used the raw scores. Because the scores are calculated in this model considering raters' self-inconsistency and item difficulty, more accurate evaluation score can be obtained by adopting the model. (2) The objective measurements adopted in the present study are selected from previous studies in the field of Second Language Acquisition (Ano, 2001; Foster, and Skehan, 1996; Iwashita, McNamara, and Elder, 2001). Muto et al (ibid) adopted the objective measurements only in terms of rhythmic feature of L2 speech, and aims at automatic evaluation of rhythmic features, but the present study attempts to investigate oral reading evaluation. (3) The present study selected the evaluation items from previous studies in English language education (Nakano, Tsutsui, Kondo, Owada, and Ueda, 2006; Yashiro, Araki, Higuchi, Yamamoto, and Komissarov, 2001; Yuan and Ellis, 2003), and then the evaluation items applicable to oral reading evaluation were selected. (4) While Muto et al (ibid) concentrated on Japanese learners of English, the present study collected the data from Asian learners of English with different first language (L1) backgrounds: Korean, Tagalog, Chinese, Malay, Thai, and Japanese, and examined the applicability of the framework proposed by Muto et al (ibid). (5) The present study adopted a short passage as the reading material because a certain length of reading in view of time is needed when raters evaluate learners' speech.

2. Method

2.1 Participants

Participants are sixteen second language (L2) learners of English with six different first language (L1) backgrounds: Thai, Malay, Tagalog, Chinese, Japanese, and Korean. They are graduate or undergraduate students who participated in an international student seminar held at Singapore in February, 2006. The average of their age is 24.1 year with 3.2 S.D.. The average length of their study of English is 13.9 year.

2.2 Recording procedure and material

The participants came to the room for recording individually, and gave their self-introductions to the interviewer, the aim of which was to reduce their tensions with the interview because some of the participants do not seem to be conditioned to the interview in English. After the self-introduction the participants read the material silently and were given the opportunity to ask the meaning of unknown words or phrases to the interviewer. If a participant did not understand the story of the reading

material, the interviewer explained it in English. Then the participants read aloud the reading the material aloud and were digital-tape recorded. It took about fifteen minutes for each participant to complete this procedure. After the recording the participants are given small gifts.

The material that the participants read out is a fable “North Wind and the Sun”. The reason why this story was chosen are: (1) It is so famous that the students at university level must know it; (2) It is used as the reading material in the NIE corpus (Deterding and Ling, 2005) which are corpus of Singapore English; (3) It is used in the phonetic description of the International Phonetic Association.

2.3 Raters and rating procedure

Ten raters whose academic backgrounds are Applied Linguistic or related area participated in this study. They attended three rater trainings where they consulted the learner description of Common European Framework of References (CEFR) (Council of Europe, 2001), watched the video provided by Language Policy Division Eurocentres (2003), and discussed the characteristics of learners’ speech of six levels proposed by CEFR. The aim of this procedure is to lead the raters to establish images of learners of six levels. After the three rater trainings the raters evaluated the learners’ speech on the website individually.

Table 2.1: Evaluation Item

Voice volume	Intonation
Voice pitch	Speech rate
Vowel quality	Frequency of pause
Consonant quality	Position of pause
Epenthesis	Approachableness
Deletion	Tightness
Rhythm	Foreign accentedness

The evaluation items were selected from Yashiro et al (2003), and the items were scrutinized in Nakano et al (2006). In the present study the evaluation items applicable to the reading evaluation were selected (Table 2.1). In each evaluation item 6-point Likert scale were adopted according to the six level proposed by CEFR.

2.4 Objective measurements

Five indices were selected from Muto et al (2003) and the previous studies in English Language Education (Ano, 2001; Foster, and Skehan, 1996; Iwashita, et al, 2001). Table 2.2 shows the indices and the description.

Table 2.2: Description of objective measurements

Index	Description
Length of utterance	The amount of time it takes for a participant to read the passage.
Number of pause	Number of pause a participants gives while reading the passage.
Number of error	Number of error a participant makes while reading the passage
Word group	Average number of syllable in a word group divided by pauses
Ratio of weak syllable length	Ratio of weak syllable time length in the passage

2.5 Analysis

Multi-regression analysis is adopted in order to investigate the relationship between the evaluation score and the objective measurements. As the criterion variable, the evaluation score, ability of examinee was used in multi-faceted Rasch model shown below. As the predictor variables the indices shown in Table 2.2 was used. In the multi-regression analysis stepwise procedure is used, and variables whose significance level of partial correlation coefficient is higher than .05 will be deleted.

$$\log(P_{nmijk}/P_{nmijk}\text{-}1)=B_n - A_m - D_i - C_j - F_k$$

where

B_n = ability of examinee n

A_m = difficulty of task m

D_i = difficulty of skill item i

C_j = severity of judge j

F_k = difficulty of category k relative to category k - 1

P_{nmijk} = probability of rating of k under these circumstances

P_{nmijk}-1 = probability of rating of k – 1

3. Results

Table 2.3 shows correlation coefficients between the objective measurements and the evaluation score. The stepwise multi-regression analysis found two dominant predictors, Length of utterance and Ratio of weak syllable. The partial correlation coefficients are -.71 and -.32, respectively. The determination coefficients R^2 is .70 ($F(2, 14) = 19.24$, $p < .01$).

Table 2.3: Correlation coefficients between the objective measurements and the evaluation score

	1	2	3	4	5	6
1	1	-.49	.53	-.52	-.58	-.80
2		1	-.97	.18	.55	.64
3			1	-.19	-.43	-.67
4				1	.40	.28
5					1	.69
6						1

1. Evaluation score
2. Number of pause
3. Word group
4. Ratio of weak syllable
5. Number of error
6. Length of utterance

4. Discussion and conclusion

The results of multi-regression analysis indicate that speech rate is a dominant predictor of the evaluation score. This is the same result as in Nakano et al (2006) and Muto et al (2003). This fact suggests not only the validity of the present study but also the applicability of the data of reading aloud a passage in L2 speech evaluation. The difference between Nakano et al, Muto et al, and the present study is the data format. Nakano et al adopted the data of unprepared speech; Muto et al, the data of reading aloud short sentences; and the present study, the data of reading aloud a passage by L2 learners. In those three studies speech rate is the most dominant predictor of the evaluation scores, however. Furthermore, the present study found the index of isochrony (Ratio of weak syllable) to be the significant predictor. The index of isochrony adopted in the present study, Ratio of weak syllable, is a tentative measurement. There is a plenty room for investigating the index of isochrony, which makes the predictability of the L2 speech evaluation accurate.

References

Ano, K. (2001). Koukousei eigo gakushusha no hatsuwa ni okeru ryuchosa to seikakusa no kankei. [The relationship between fluency and accuracy in the utterances of high school students of English]. *STEP Bulletin*, 13, 39-49.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: CUP.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18, 3, 299-323.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, 51,3, 401-436.

Language Policy Division, Eurocentres, (2003). *Samples of oral production illustrating for English, the levels of the CEF of reference for languages*. Eurocentres.

Muto, M., Sagisaka, Y., Naito, T., Maeki, D., Kondo, A., & Shirai, K. (2003). Corpus-based modeling of naturalness estimation in timing control for non-native speech. *Proceedings of Eurospeech* 2003. 401-404.

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