Fundamental Research on Automatic Speech Evaluation of L2 English

Michiko Nakano¹, Yusuke Kondo², Eiichiro Tsutsui³

¹Faculty of Education and Integrated Arts and Sciences, Waseda University, ²Language Education Center, Ritsumeikan University, ³International Communication Center, Hiroshima International University

nakanom@waseda.jp, ykondo@fc.ritsumei.ac.jp, tsutsui@ic.hirokoku-u.ac.jp

Abstract
The final goal of this study is to build an automatic speech evaluation system for Second Language English. Applying our automatic speech evaluation system, we aim to replicate by computer the subjective evaluation of speech characteristics done by trained language teachers, using objective measures. Therefore, to detect predictors of the subjective evaluation scores which are latent in the objective measures, we investigated the relationship between the evaluation scores and the objective measures. We report here the procedure of the subjective evaluation to arrive at high reliability, and satisfactory correlations between human ratings and objective measures. In order to replicate subjective evaluation by using objective measures, we used multi-regression analysis to obtain linear expression. The correlation between the observed values and the predicted values is .69.

Keywords
Second Language speech, Automatic evaluation, Multi-regression analysis

Introduction
In the second language (L2) assessment, oral interview tests are often employed as in ACTFL OPI, STEP TESTS, and various versions of Cambridge Proficiency Tests, etc.; they are asked to discuss general topics, or they are asked to describe pictures and role-play in prescribed situations. In these tests, their oral performance is assessed manually by trained raters based on the respective criteria of proficiency standards. The process of human judgments takes time, involving tedious process of listening to the recorded videos several times, in order to arrive at good inter-rater agreement.

In this pilot study, it is hoped that we can implement automatic oral performance by L2 users of Englishes in the near future; namely, computer can assess oral performance in which the objective measures should replicate subjective evaluations made by experienced English teachers. In Muto, Sagisaka, Naito, Maeki, Kondo, and Shirai (2003), they focused on temporal features as predictors of subjective evaluations. Muto et al. (2003) investigated the relationship between subjective evaluation and objective measurements in terms of rhythmic control in order to establish an automatic evaluation system for L2 speech. The present study is based on the framework of Muto et al (2003). To initiate the matter, we collected the data of 101 reading-aloud speeches from Asian learners of English whose language backgrounds were Japanese, Korean, Chinese, Filipino, Thai, and Malay.

All the data were assessed by experienced language teachers. In the process of the evaluation we conducted rater training based on Common European Framework of References (CEFR), and scrutinized the evaluation items based on the raters’ feedback, and counted out unreliable raters based on statistics. Adopting the evaluation scores which we obtained through this process, we investigated the relationship between subjective evaluation and objective measures.

1 Data
1.1 Participants
The participants were 101 Asian learners of English whose language backgrounds were Japanese, Korean, Chinese, Filipino, Thai, Khmer, Malay, Vietnamese, and Indonesian.

Table 1.1: Key information of the participants

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>23.46</td>
<td>4.42</td>
<td>18-38</td>
</tr>
<tr>
<td>Study of English (year)</td>
<td>11.88</td>
<td>5.41</td>
<td>2-31</td>
</tr>
</tbody>
</table>

They were either undergraduate or graduate students. Table 1.1 shows the average, standard deviation, and range of their age and study of English.
1.2 Text and recording procedure

The reading text that the participants read was a fable from Aesop, “The North Wind and the Sun” (Appendix A), which is so famous that the students at university level must know it. This passage is also used in the NIE corpus (Deterding and Ling, 2005), and is used in the phonetic description of the International Phonetic Association.

The participants came to a recording studio individually, introduced themselves in English, and were instructed how to read the passage in English. Their English performance was recorded using a digital recorder, Roland R-09 and a condenser microphone, SONY ECM-MS957. It took around fifteen minutes for each participant to complete this procedure.

1.3 Rating Procedure

Raters who joined this experiment were six Japanese language teachers of English whose academic background is Applied Linguistics or related areas. The raters evaluated the recorded speeches on the website individually. As mentioned in later sections, all the raters participated in the raters’ workshop, and some unreliable raters were deleted in the analysis based on statistics.

2 Subjective evaluation

2.1 Item selection

The evaluation items are originally from Yashiro, Araki, Higuchi, Yamamoto, and Komissarov (2001). Their items are to abstract away not only linguistic, but also para-linguistic characteristics. From their items, we chose several items suitable to evaluate recorded speech based on the discussion of raters’ workshop. In the selection of the evaluation items, we analyzed the items based on Item Response Theory (IRT) using statistical software, FACETS (Linacre, 2006). In this model, item properties and trait level can be separately estimated. The model is depicted in the equation (1). To scrutinize items, the scores of InfitMS, the misfit of items was adopted. The score of InfitMS is an index of the difference between observed scores and expected scores (McNamara, 1996: 169-176). According to the criteria in Kondo-Brown (2002), we considered items with above or below ±2 SD of InfitMS as the items unsuitable to the construct we try to measure. Table 2.1 shows the items used in the present study. Our raters used these fourteen items to evaluate the learners’ speeches. We adopted 6-point Likert scale, because as mentioned later, our raters received the rater training according to CEFR which describes learners of six levels.

\[
\log \left( \frac{P_{nmijk}}{P_{nmijk-1}} \right) = B_n - A_m - D_i - C_j - F_k
\]  

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 \(k - 1\)

2.2 Rater training

We conducted our rater training according to CEFR. CEFR is widely used guidelines on learning, teaching, and assessing L2 and describes six levels of proficiency with descriptors. Reception, production, and interaction modes refer to both spoken and written modes; therefore there six categories in CEFR, and the descriptors are represent Can-do statements for the learners’ linguistic activities. In addition to the descriptors in global scales, such as spoken interaction, and written production, CEFR presents the descriptors in local scales such as phonological control and grammatical accuracy. Table 2.2 shows an example of descriptors in CEFR.

<table>
<thead>
<tr>
<th>PHONOLOGICAL CONTROL</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
</tr>
<tr>
<td>Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.</td>
</tr>
<tr>
<td>C1</td>
</tr>
<tr>
<td>As in C1</td>
</tr>
<tr>
<td>B2</td>
</tr>
<tr>
<td>Has acquired a clear, natural, pronunciation and intonation.</td>
</tr>
<tr>
<td>B1</td>
</tr>
<tr>
<td>Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.</td>
</tr>
<tr>
<td>A2</td>
</tr>
<tr>
<td>Pronunciation is generally clear enough to be understood despite a noticeable foreign accent, but conversational partners will need to ask for repetition from time to time.</td>
</tr>
<tr>
<td>A1</td>
</tr>
<tr>
<td>Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by native speakers used to dealing with speakers of his/her language group.</td>
</tr>
</tbody>
</table>

(Council of Europe, 2003:117)

In our rater training, the raters watched the videos provided by Council of Europe and discussed the characteristics of learners grouped into six levels. The videos depicted the learners divided into six levels according to CEFR. After that, raters discussed the characteristics of learner language in each level. This activity might lead our raters to establish the images
of the learners of six levels.

To investigate the effect of the rater training, Nakano, Kondo, Tsubaki, and Sagisaka (2008) examined the reliability improvement of the evaluation score adopting Generalizability Theory (G Theory). They compared the evaluation scores before and after their rater training. The raters evaluated recorded unprepared speech by Asian learners of English. The remarkable difference before and after our rater training is in the estimated variances of the items. The estimated variance of items after the training is about one-sixth of that of items before the training. This suggests that the items (rating criteria) before the training differ much more in average difficulty than the items after the training. Furthermore we counted out some unreliable raters according to the same procedure as that of item selection as presented section 2.1.

2.3 Evaluation reliability in the present study

In the present study, the six raters who received the rater training joined this experiment. To examine the reliability of the evaluation in the present study, we analyzed the evaluation scores based on G Theory and compared the index of dependability of the evaluation scores in the present study with that of the evaluation scores in Nakano et al (2008). The data from Nakano et al were the evaluation score by the raters before receiving their rater training.

Figure 2.1: The Shift of the Index of Dependability

Evaluation conditions were simulated where one to ten rater(s) evaluated examinees using ten items. Figure 2.1 shows shift of the index of dependability: y axis is the index of dependability, and x axis, number of raters. The ash line is the shift of the index of dependability of the evaluation scores from Nakano et al (2008), and the black line, that of the present study. With the acceptance that this index is the analogue of reliability coefficient, the minimum value is .8 for a reliable evaluation. In the evaluation reliability of the present study, the index exceeds the standard in the case of the six raters. Furthermore, in the case of the evaluation of reading aloud speech, we can obtain more reliable evaluation score if six or more trained raters joined the evaluation than in the case of the evaluation before the rater training in Nakano et al (2008).

3 Objective measures

3.1 Overview

First of all, we selected objective measures which could be possible predictors of the evaluation scores. We conducted several pilot studies where the correlations were examined between the objective measures, and between objective measures and the subjective evaluation scores. Judging from the correlation coefficients and partial correlation coefficients in multiregression analysis, we selected the objective measures as the predictors of the scores.

3.2 Speech timing control characteristics


- Number and duration of filled pause: Insertion of mm, aan, aam, ehh
- Number and duration of silent pause (100 ms): Counted every 10 ms from 10 ms to 400 ms and considered the correlation coefficients between each pause (10 ms, 20 ms, 30 ms …) and the evaluation scores.
- Mean length of run: Average number of syllable between immediately after a silent and the end of utterance
- Pruned syllable per second
- Ratio of unaccented syllable to accented syllable

As for the index of speech rate, we examined the relationships between the evaluation scores and the three indices of speech rate: syllable per second, articulation rate, and pruned syllable per second. Syllable per second is total number of syllables including self-correction, self-repetition, and filled pause divided by the total number of seconds. Articulation rate is Total number of syllables including self-correction, self-repetition, and filled pause divided by the total number of seconds excluding silent pause. Pruned syllable per second is total number of syllables excluding self-correction, self-repetition, and filled pause divided by the total number of seconds excluding silent pause. This is based on Munro and Thomson (2004). Judging from the correlation with the evaluation score, we adopted pruned syllable per second as the index of rhythm. Pruned syllable per second is operationalized as follows:
\[ S = \frac{T - E}{TD} \]

\( S \) is the index of speech rate, \( T \) is the total number of syllable a learner uttered, \( E \) is the total number of syllable unneeded (e.g. repetitions, fillers, and false starts), and \( TD \) is the total time duration.

Ratio of unaccented syllable to accented syllable is operationalized as follows:

\[ R = \frac{A}{U} \]

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

As Table 3.1 shows, Kondo et al (2007) found two good predictors of the score: pruned syllable per second and ratio of unaccented syllable to accented syllable. Syllable per second obtained the highest correlation with the score. Ratio of unaccented syllable to accented one, the index of rhythm correlated with the score independently from other variables. They found the high correlation coefficients in number of silent pause and mean length of run, but they also correlates with other variables.

Table 3.2: Examples of pause categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentential pause</td>
<td>...when a traveler came along wrapped in a warm cloak. &lt;P&gt; They agreed that the...</td>
</tr>
<tr>
<td></td>
<td>...the North Wind was obliged to confess that &lt;P&gt; the Sun was the stronger of the two.</td>
</tr>
<tr>
<td>Phrasal pause</td>
<td>And at last &lt;P&gt; the North Wind gave up the attempt.</td>
</tr>
<tr>
<td>Within-phrase pause</td>
<td>The &lt;P&gt; North Wind and the Sun were disputing...</td>
</tr>
<tr>
<td></td>
<td>...the Sun shone out &lt;P&gt; warmly and immediately the traveler...</td>
</tr>
</tbody>
</table>

Table 3.3: Correlations between score and measures

<table>
<thead>
<tr>
<th>ES</th>
<th>SR</th>
<th>SP</th>
<th>PP</th>
<th>WP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES</td>
<td>1</td>
<td>-.83</td>
<td>-.15</td>
<td>-.64</td>
</tr>
<tr>
<td>SR</td>
<td>1</td>
<td>-.34</td>
<td>-.75</td>
<td>-.54</td>
</tr>
<tr>
<td>SP</td>
<td>1</td>
<td>-.32</td>
<td>-.19</td>
<td></td>
</tr>
<tr>
<td>PP</td>
<td>1</td>
<td>.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WP</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: ES, SR, RH, SP, PP, and WP stand for evaluation score, index of speech rate, and rhythm, average number of sentential pause, phrasal pause, and within-phrase pause respectively.

3.4 Vowel discrimination

Kitagawa, Kondo, and Nakano (2007) examined the relationships between the vowel discrimination rates and the scores. This is only for Japanese participants. They acoustically examined the vowel quality produced by Japanese participants, and investigated the relationship between the vowel classification rates and the evaluation score adopting discriminant analysis. As Table 3.4 shows, some tendency was found,
but the vowel classification rates could not be a good predictor of the score.

Moreover, Kitagawa and Kondo (2008) investigated the differences of fundamental frequency, intensity, and time duration of reduced vowel /ə/ among the three levels of learners, and examined the relationship between the measures and the levels, but they could not find a good predictor among the measures.

Table 3.4: Correct classification ratio of vowels

<table>
<thead>
<tr>
<th></th>
<th>High</th>
<th>Middle</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>/i/ vs. /˘i/</td>
<td>86.7</td>
<td>51.4</td>
<td>43.3</td>
</tr>
<tr>
<td>/u/ vs. /˘u/</td>
<td>63.3</td>
<td>63.3</td>
<td>56.7</td>
</tr>
<tr>
<td>/˘i/ vs. /˘u/</td>
<td>62.9</td>
<td>53.7</td>
<td>55.9</td>
</tr>
</tbody>
</table>

4 Correlation study

4.1 Method

Judging from the results in the pilot studies, in the present study, we chose two predictors of the score: syllable per second, namely index of speech rate and ratio of unaccented syllable to accented syllable, namely index of rhythm.

To investigate the predictability of the evaluation score by the objective measures, indices of speech rate and rhythm, multi-regression analysis was adopted. The criterion variable is the evaluation score, and the predictor variables, the indices of speech rate and rhythm.

4.2 Results

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 4.1 is the scatter graph of the observed and predicted value: y-axis is the observed value and x-axis, the predicted value.

Figure 4.1: The Observed and Predicted Score

5 Summary and discussion

In this study we attempt to replicate subjective evaluation by using objective measures. To obtain linear expression, we used multi-regression analysis. Although we can find some outliers in our data as Figure 4.1 shows, we obtained high multiple correlation coefficient (.69) in this study. Considering the difference between previous studies and the present one, we assume that this high coefficient is attributable to the three factors: 1) the reading text which is long enough for raters to detect learners’ characteristics of speech, 2) the rater training with clear criteria, CEFR, and 3) rater selection based on IRT.

The final goal of this study is to build an automatic speech evaluation system of L2 English. 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. We need to examine the outliers in detail.

Appendix A: Reading text

The North Wind and the Sun were disputing which was the stronger when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.

Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveler fold his cloak around him; and at last the North Wind gave up the attempt.

Then the Sun shone out warmly, and immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

References


