Waseda University

The development of automatic speech evaluation system for learners of English

Dissertation

submitted in the partial satisfaction of the requirements for the degree of Doctor of Education

Yusuke Kondo

2010

Acknowledgements

The studies in this dissertation were conducted from 2005 to 2009. These studies reflect the relationships with many inspiring people whom I have met since the beginning of my work. Many people have assisted me in the preparation of these studies, but, of course, I alone am responsible for any mistakes and shortcomings. I would like to express my deep gratitude to all those who gave me the possibility to complete this dissertation.

First and foremost, I am heartily thankful to my supervisor, Michiko Nakano whose unfailing encouragement, immeasurable support, sage advices, and critical comments from the beginning to the concluding level enabled me to complete the dissertation. I am also thankful to my co-supervisors, Yoshinori Sagisaka, Testuo Harada, and Yasuyo Sawaki for their insightful comments and suggestions.

I also wish to extend my special thanks to my co-researchers, Kazuharu Owada, Norifumi Ueda, and Eiichiro Tsutsui. They helped to establish a very important basis for the primary investigation of this study.

I offer my regards and blessings to all the students and the teachers who participated in these studies as an informant or a rater. Without their participation, I could not finish my dissertation.

Last but not least, I wish to express my gratitude to my wife who has been always there cheering me up and stood by me through the good times and bad.

Table of contents

List of Tables

List of Figures

1 Introduction

1.1 Statement of the problems

Speaking is one of the essential skills in the attainment of second language (L2) learning, and speaking skills are also important objectives in L2 assessments. 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 sort of speaking tests, test designers discuss and determine a set of evaluation criteria and the procedure of rating, and raters receive some training to arrive at good inter-rater agreement. After the test, the raters sometimes watch the video or listen to the recorded speech of the examinees', and the evaluation scores given by the raters are analyzed based on some statistical model. The process of this sort of test takes time; especially the implementation of speaking tests, such as an interview or a picture description task. Therefore, though the importance of the speaking tests is generally recognized, in many cases, a speaking test is not adopted as an achievement test or a placement test. The reasons often include the costs of a speaking test: the time for the implementation of the test and the evaluation by human raters. 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.

Computerized assessment is one of the solutions to reduce the cost related to language testing. As Jamieson (2005) mentioned, the recent development of computer technology effects improvements in language testing. The first change of language testing by the introduction of computer was such automated assessment as those in which examinees answer

multiple choice or fill-in-the-blank items. In this sort of computerized test, the marking was done by the computer, so that the some of the cost reduction was achieved. In the next step, as a test that cannot be conducted in the format of paper-and-pencil tests, computer adaptive tests are now available to measure receptive skills, listening and reading. However, though some implementations are found such as Versant, English Communication Assessment Profile (E-CAP), and TOEFL Online Practice, as for the assessment of the speaking and the writing skill, which requires human judgments, the automatic evaluation system is now in the process of being developed.

The construction of automatic L2 speech evaluation system requires the precise measurement of learners' speech characteristics, the careful consideration in the procedure, and the analysis of the evaluation, because human rating is predicted by using learners' speech characteristics in the current approaches in the automatic evaluation system. The measurement and the evaluation of L2 speech were done in several areas of research, such as Speech Science, Educational Measurement, and Applied Linguistics. Accurate acoustic measurement of learners' speech characteristics requires technologies of speech recognition studied in Speech Engineering, and mathematical models used in Educational Measurement is needed to examine the reliability in the evaluation scores given by human raters. Furthermore, the procedure of test and evaluation should be designed in the viewpoint of English language education. However, until recently neither researchers in Speech Engineering reflected on the findings in Applied Linguistics in their research designs, nor researchers in Applied Linguistics paid enough attention to measurement models to their analyses of the evaluation. The collaborative research on the development of automatic speech evaluation system has only begun (e.g. Xi, Zechner, and Williamson, 2008; Bernstein, 1999).

In the development of automatic speech evaluation system, large quantities of speech data

are required, because the automatic L2 speech evaluation will be based on the analysis of the learners' speech characteristics. Based on the analyzed speech data, the system predicts the human judgments and gives the examinees the evaluation scores. Recently, L2 speech corpus and automatic L2 speech evaluation system have been studied in Speech Science. However, there are some problems to be solved. Firstly, in these studies, raters who had received no training evaluated the learners' speech using an evaluation item. In some cases, raters received training for rating, but no criteria were referred to in the rater training, and in other cases a single rater evaluated learners' performance. Furthermore, the text that the learners read out was so short that the rater appeared to be unable to catch the learners' speech characteristics realized in the speech. This is an unusual situation in language testing of L2. The evaluation scores will be analyzed based on a certain statistical model, and unreliable items and raters will be excluded. Since the introduction of Generalizability Theory and Multi-faceted Rasch Analysis into the field of language testing, applying these techniques, L2 performance evaluations have been analyzed to obtain reliable evaluation scores (e.g. Kunnan, 1992; Akiyama, 2001; Kondo-Brown, 2002; Bonk and Ockey, 2003; and Kozaki, 2004).

In the analysis of performance assessment, the three facets of examinees, raters, and evaluation items need to be analyzed. To examine the reliability of the raters and the items based on Classical Test Theory (CTT), usually one of the facets, the raters, is compressed by calculating the average or the median of scores among raters. Therefore, the analysis would be done as if only one rater evaluated examinees. In this case, it is impossible to examine the reliability of raters and items at the same time. Because the index of reliability often used in CTT is supposed to consist of two facets, items and examinees, this index cannot be applied to the performance assessment. The reliability estimation in CTT has several limitations: The CTT estimations detect only one source of error in a single analysis, do not deal with systematic error and random error separately, and estimates reliability and standard

error measurements are equal for scores at all levels (Bachman, 2004: 174).

In the evaluation of performance assessment, raters are usually asked to evaluate examinee performance, using several evaluation items. In each evaluation item, raters usually rate the examinees performance as 1 (poor) or 5 (very good) in terms of the aspect of the performance which the evaluation item depicts. In the situation where the raters receive neither training nor instruction about the evaluation, the scoring will be left to the judgment of the individual raters. This could be one of the factors lowering the reliability of the assessment.

Another problem to be solved is the eligibility of raters. In the automatic evaluation system, the scores are predicted based on human judgment. In almost all of the automatic speech evaluation system, the human ratings were done by the native speakers of the target language (e.g. Neumayer, Franco, Digalakis, and Weintraub, 2000; Cucchiarini, Strik, and Boves, 2000a; Cucchiarini, Strik, and Boves, 2000b; Cucchiarini, Strik, and Boves, 2002; Zechner, Higgins, Xi, and Williamson, 2009; and de Wet, Van der Walt, and Niesler, 2009). In some cases, the scores given to examinees are calculated, based on the differences between the native speakers of the target language and the examinees. This does not suit to the situation of English language education in countries where English is learnt as a second or foreign language. L2 learners are taught and evaluated by their teachers who are users of that language. About eighty per cent of English language teachers in the world are L2 users (Canagarajah, 1999). Furthermore, McKay (2002) insisted that from the view point of Worled Egnlishes, L2 users were be more eligible than the native speakers of English, and Kim (2009) demonstrated that both native speakers and non-native speakers were equally reliable in the analysis of L2 learners' performance assessment. Hence, the automatic evaluation system to be constructed in this situation should predict the evaluation by the teachers who are the L2 users.

1.2 Purpose of the study

The purpose of this study is to build an automatic L2 speech evaluation system which predicts evaluation scores given by experienced teachers who are L2 users. Several steps are required to accomplish the purpose. Firstly, a criterion of the evaluation is examined. The criterion to be examined in this study is Common European Framework of Reference for languages (CEFR). CEFR is a widely used guideline on learning, teaching, and assessing L2 proficiency and describes six levels of learners with descriptors. A companion piece of CEFR, European Language Portfolio (ELP), is a self-assessment tool of language proficiency developed in parallel with CEFR. A great number of pilot projects were conducted to develop ELP (Little, 2002). However, these are conducted in the European context. Before adopting the criterion, the applicability is examined in the Asian context by using ELP. Secondly, the eligibility of L2 users as raters in L2 performance assessment is examined: rater training for L2 speech evaluation is conducted according to CEFR, and the effect of the training is examined based on two statistic models, Generalizability Theory (G-Theory) and Multifaceted Rasch Analysis (MFRA). Thirdly, the rating procedure, the reliability of raters and evaluation items are examined through two types of L2 speech performance: spontaneous speech and read-aloud speech. The predictability of the evaluation scores by speech characteristics of learner performance data is investigated in these two types of speech. Fourthly, based on the results of the studies mentioned above, an automatic L2 speech evaluation system is constructed, and its reliability is examined.

1.3 Outline of the study

This dissertation consists of nine chapters. In Chapter 2, the related works to the present study are reviewed: the criterion of the evaluation, CEFR, statistic models used in the analysis of the evaluation, G-Theory and MFRA, and studies on speech characteristics of L2 learners'.

In Chapter 3, the examination of the applicability of CEFR and ELP is reported. Then, in Chapter 4, the effects of the rater training according to CEFR are investigated. Chapter 5 and 6 examine the predictability of the evaluation scores by speech characteristics realized in the two types of speech. In Chapter 7 the details of the speech database are described to construct the automatic L2 speech evaluation system. In Chapter 8, the construction of the automatic L2 speech evaluation system is described, and its reliability is examined. Finally, in Chapter 9, I summarize and conclude the dissertation, and point out the limitations of this study.

2 Background study

2.1 Overview

This chapter reviews the works related to the present study. Firstly, the criterion to be adopted in this study is outlined. In this study, Common European Framework of Reference was adopted as the criterion in the second language (L2) speech performance assessment. The practical reasons were given for the adoption of CEFR, compared to other criteria for L2 assessment. Secondly, the eligibility of the raters in L2 performance is discussed from the view point of World Englishes. Thirdly, the models of statistical analyses to be used, Generalizability Theory (G-Theory), Item Response Theory (IRT), and Neural Test Theory (NTT) are described. The advantages of these statistical models over Classical Test Theory (CTT) are mentioned. Lastly, studies on speech characteristics of L2 learners' are reviewed. Studies in several fields have investigated the relationship between human rating on L2 speech and speech characteristics of L2 learners'. Their shortcomings are pointed out.

2.2 Common European Framework of Reference and European Language Portfolio

Attempts have been made to describe the development of L2 learners' proficiency, which is essential in composing a test, developing a language learning curriculum, and self-evaluating language ability. However, as North and Schneider (1998) indicate, there is no language proficiency model that is empirically and theoretically valid, and the examination of validity of proficiency scales or descriptors involves extensive research. Therefore, as of this moment, we cannot obtain proficiency scales or descriptors based on an established language proficiency model.

An early study of the description of the development of L2 learners' proficiency, Foreign Service Institute (FSI) scales were developed in 1950s. FSI comes down to American

Council on the Teaching of Foreign Language (ACTFL) Proficiency Guidelines (American Council for the Teaching of Foreign Languages, 1999). In ACTFL, learners are evaluated with ten levels in four language skills: listening, speaking, reading, and writing. In the evaluation methods provided by ACTFL, Oral Proficiency Interview (OPI) which take fifteen minutes to twenty five minutes, interviewers control the levels of questions to examinees and tasks for examinees to accomplish. Standing on the theoretical foundation of OPI, Standard Speaking Test (SST) was developed by ALC Press to meet the needs of Japanese learners of English (ALC Press, 2006). However, these two tests have been criticized for the low validity and reliability (e.g. Lee and Musumeci, 1988; Salaberry, 2000). Lee and Musumeci (1998) pointed out that the tasks in OPI and SST were not hierarchically arranged: the skills and the ability required in the tasks of higher levels do not postulate those required in the task of lower levels. Furthermore, Salaberry (2000) noticed that improvement had not been occurred in the ACTFL Tester Training Manual published in 1999 from the previous manual published in 1986.

Another framework of foreign or second language learning related to ACTFL is Canadian Language Benchmarks (CLB: Centre for Canadian Language Benchmarks, 2000). The purposes of CLB are to provide learners with indices to be used in the self-evaluation of L2 ability, and provide a commonly understood framework for language programs in Canada. In CLB, in terms of four language skills: listening, speaking, reading, and writing, learners are divided into twelve levels. In each level, in addition to can-do statements, typical examples of tasks and texts, performance indicators, and strategies to be taught are provided. However, CLB does not include descriptions of discrete knowledge and skills (e.g. pronunciation, grammar, and vocabulary).

The European counterpart of ACTFL is CEFR (Council of Europe, 2001). CEFR is a widely used guideline on learning, teaching, and assessing L2 and describes six levels of

learners with descriptors. In reception, production, and interaction, the descriptors of language proficiency in relation to learners' activities are listed with respect to the six levels. 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. CEFR presents detailed descriptors which capture various aspects of learners' activities. The descriptors of CEFR are written, based on theories of language competence and scaled based on a theory of measurement. In CEFR, learners are initially divided into three levels; basic user, independent user, and proficient user, and then each level is divided into two levels, which makes the six levels; Breakthrough, Waystage, Threshold, Vantage, Effective Operational Proficiency, and Mastery. Each level is usually called A1, A2, B1, B2, C1, and C2 respectively. The number of the levels is largely based on the works by Trim and Wilkins (e.g. Trim, 1978). The scaling of the descriptors has been examined by a large number of researches (Council of Europe, 2001: 217-225).

In North and Schneider (1998), two projects were reported: the one is for English, and the other for French and German. The aim of the projects was to develop a scale of language proficiency in the forms of descriptors. This is a fundamental research on validation of descriptors and levels in CEFR. The projects consisted of three stages to scale the descriptors. In the first stages, descriptors were created based on models of communicative competence and language use, and then, the created descriptors were categorized into some groups, such as reception, interaction, and production. In the second stage, which they called qualitative validation, the quality and the classification of the descriptors were examined by language teachers. They held thirty two workshops attended by more than 292 teachers through these two projects for the qualitative validation of the descriptors. The purpose of this procedure was to ensure that teachers' thoughts were well represented in the pool of the descriptors. In this workshop the teachers discussed learners' performances and

sorted the descriptors into some provisional ranks. Based on the discussion and the levels of descriptors sorted by the teachers, questionnaires were composed, and the teachers evaluated learners' performances by using the questionnaires. In the third stage, the statistical analyses of the questionnaires were done based on Multifaceted Rasch Analysis (MFRA). Some descriptors were excluded based on the fit statistics produced in MFRA and Differential Item Functioning. The quality, the classification, and the levels of the descriptors were validated by comparing the results of the two projects. Although these two projects were conducted in different context of language learning: the first project was for English, and the second was for French and German, the correlation of the difficulty of the descriptors in the two projects were almost identical (r = .99), and descriptors on similar issues were adjacently aligned. North and Schneider (ibid) concluded that these results, the coherence and the consistency of the scaling of the descriptors were attributed to the facts that the descriptors were organized and selected according to the models of the communicative competence and language use, that the quality of the descriptors were examined by language teachers, and that the analyses were done based on Item Response Theory (IRT). However, they reminded us that the interpretation of the descriptors were subject to the context of language learning, and mentioned that the provision of the validated scale of language proficiency was only the first step to the establishment of an assessment framework.

A companion piece of CEFR, European Language Portfolio (ELP), which is a self-assessment tool of language proficiency developed in parallel with CEFR. The purpose of ELP is twofold: to motivate learners in language learning and to provide a record of language learning. "Can-do" statements in ELP have one-to-one correspondence with the descriptors of CEFR. For example, the counterpart of a CEFR descriptor, "Can handle very short social exchanges but is rarely able to understand enough to keep conversation going of his/her own accord, though he/she can be made to understand if the speaker will take the

trouble." is "I can handle short social exchanges and make myself understood if people help me." in ELP. More than one million copies of the "Can-do" statements in ELP were distributed among European countries and validation studies of ELP were reported from about forty European countries. Over ten thousands learners participated in each validation study (Schärer and Rapporteur, 2004). Researchers have reported the developments of curriculum and assessment for language learning based on CEFR and ELP (Morrow, 2004).

One of the examples is Hasselgreen (2005), which investigated the applicability of CEFR and ELP in the context of the assessment of young language learners in Nordic/Baltic countries. In Hasselgreen (ibid), two projects were reported. Since the descriptors and the levels in CEFR are designed for adult language learners, their applicability should be examined for young language learners. In the first project, the descriptors and the levels in CEFR were examined to adapt them for assessing the young language learners. Sixteen teachers selected appropriate levels in CEFR to describe their learners and modified the expression in the descriptors in CEFR and the "Can-do" statements in ELP to be suited to their learners. After that, 259 learners in Nordic/Baltic countries self-evaluated their language ability by using the modified "Can-do" statements. Apparently unsuitable descriptors and "Can-do" statements were adjusted or excluded through this procedure. In the second project, the tests of reading and writing skills were developed, which were the national project to report to schools, parents, and authorities on the language ability of the young language learners in Norway. All test items were designed to correspond to the "Can-do" statements. In the evaluation of writing skills, teachers received an extensive rater training based on various descriptions in CEFR before the rating. The analyses were done, based on IRT, and missfitting items were excluded, and high reliability was obtained in all the tests. Based on the results of these two projects, Hasselgreen (ibid) concluded that though we had to make consideration of the characteristics of young language learners, the levels and

the “Can-do” statements in CEFR and ELP can be applicable for the assessment of young language learners, preserving the integrity of CEFR levels. However, it was mentioned that the descriptors in CEFR and the “Can-do” statements and ELP were not sufficient to assess the everyday classroom performance by the young language learners. Furthermore, Hasselgreen (ibid) pointed out that teachers would need training for language assessment in this sort of project. This study implies the need for modification of the descriptors and the levels in CEFR and the “Can-do” statements and ELP according to the context of language learning.

It is possible to compare our learners with others if our rating procedures are implemented in relation to CEFR, which is one of the aims of CEFR (Council of Europe, 2001: 21). The manual (Council of Europe, 2003) is provided for relating examinations to CEFR, and a documented video (North and Hughes, 2003) is also available, which describes the performances of Swiss adult learners of English calibrated to CEFR levels. Council of Europe (2005) provided a reference supplement to the manual for statistic analyses, such as factor analysis, multidimensional analysis, and MFRA.

2.3 Raters, rating, and rater training

From the view point of World Englishes, English users are considered to be more eligible as educators than the native speakers of English (McKay, 2002). Now English is an international language that serves communities of businessmen and researchers all over the world. English “provides for effective communication, but at the same time it establishes the status and stability of the institutional conventions which defines these international activities.” New Englishes are locally developed in such community. The native speakers of English, British or American are irrelevant to such Englishes (Widdowson, 2003:40). Widdowson implies that learners of English have various purposes of learning English; to be a

member of the native speakers' community is one of their purposes of English learning. To acquire the competence of the native speakers of English is one of the final goals of their English learning. The final goal of vast majority of learners of English is to be a member of international communities where English is used as a communication tool (McKay, ibid). Against this background, although the quality of communication and standards of intelligibility are not assured if we fail to preserve standard (if Englishes used in the world are not mutually intelligible, the purpose of learning English disappears), English users are more eligible as educators than the native speakers of English, because English users are more knowledgeable in English learning in their community, which is no longer relevant to the native speakers of English.

Norcini and Shea (1997) mentioned, in the context of standard settings, that the most important factor in developing a credible standard is qualified standard setters. The same can be said on L2 performance assessment. Raters must be knowledgeable in their evaluation and their examinees, and particularly must be certificated. Furthermore, in L2 performance assessment, they must understand the context of learning the target language. The eligibility of raters is one of the issues to be considered in L2 performance assessment, because the property of raters, such as severity and consistency, might be influenced by their experience and language background. For these reasons, experienced Japanese language teachers were chosen as raters in the present study, because they are conversant with Asian learners of English and with the context of English language education in a situation where English is learnt as a foreign language. In addition, the rating by non-native language teachers of English is fairly realistic for Asian learners of English. According to Canagarajah (1999), eighty per cent of English language teachers in the world are non-native speakers of English. Japanese secondary education follows the similar pattern: there are only about 4700 English teachers who are the native speakers of English. That means that

one native English teacher has 1200 students in the secondary education (Takanashi, 2009: 184). In the Takanashi's data, only the students in public school were included. If the number of the students in private schools is added, that of students per one native English teacher will explode. The situation indicates that, generally speaking, learners of English have the slightest chance to be evaluated by the native speakers of English.

However, the eligibility of L2 users as a rater in L2 performance evaluation is questionable. Kim (2009) gave an answer to this question. She investigated the differences of rating behaviors between Korean teachers and Canadian teachers in evaluation of an oral proficiency test administered to ten Korean students at a university. The evaluations were analyzed based on MFRA. The index of self-consistency in the evaluation adopted in this study were fit statistics, proportions of large standard residuals between observed and expected scores, and a single rater-rest of the raters correlation. The results revealed that in the severity and the self-consistency, there was little difference between non-native speakers of English and native speakers of English. The two groups of teachers showed the same pattern in the severity of the evaluation, and all teachers fell into the acceptable range of the self-consistency. Kim (2009), according to the results, concluded that non-native speakers of English were able to function as reliable raters in L2 performance evaluation, with the caveat that the results of the study might not be applied to other L2 performance evaluation, because only Canadian and Korean teachers were included as the raters.

Assessments of human performance require a number of raters, because no one evaluation can be definitive. A number of raters will be needed to obtain valid evaluation of human performance. Raters do not always agree, however. Therefore, rater training is usually conducted in order to achieve certain agreement among raters. As recent studies on L2 performance evaluation revealed (Lunz, Wright, and Linacre, 1990; Weigle, 1998), rater training is not capable of letting raters to achieve the same level of severity, but to make the

raters self-consistent. As shown in Weigle (1998), rater variability cannot be eliminated, but extreme differences can be reduced. However, because the difference of the severity among raters can be modeled in MFRA to some extent (McNamara, 1996: 233), the reduction of the variability in raters' severity is not a main purpose of rater training, but the focal point of rater training is to let raters to be internally consistent in their evaluation.

2.4 Generalizability Theory and Multifaceted Rasch Analysis in L2 performance assessment

The automatic speech evaluation system to be constructed here is the system which predicts the evaluations given by human raters by using objective measures of speech characteristics. This assessment model requires a reliable criterion variable. The predictor variables are winnowed down in terms of their predictability of the criterion variables. Hence, the examination of the evaluations by human raters is an essential part in the preliminary stage of examining predictor variables by multiple regression analysis.

In this study L2 speech evaluations are analyzed based on G-Theory and MFRA. These two techniques work in a mutually complementary manner in the analysis of performance assessment. While G-Theory detects the source of error in each facet: rater, item, and examinee, and on the other hand, MFRA provides information on specific raters, items, and examinees that reduce the reliability of the performance assessment. These two approaches to the analysis of performance assessment were often adopted by studies on L2 performance (Bachman, Lynch and Mason, 1995; Lumly and McNamara, 1995; Weigle, 1995; Kozaki, 2004; Bonk and Ockey, 2003; Kondo-Brown, 2002).

Bachman, Lynch and Mason (ibid) and Lumley and McNamara (1995) adopted these two techniques to analyze the performance assessment of L2 speaking ability. Bachman et al. (ibid) used these two techniques to analyze the data of a foreign language (Spanish) performance assessment for the placement of students at University of California, and

investigated the reliability of the assessment. They mentioned that test users must have adopted some models to detect multiple sources of measurement errors, and G-theory and MFRA were not anti-theoretical model of measurement, but they give us complementary information in the analysis of performance assessment. Lumley and McNamara (ibid), using these two approaches, analyzed a test of communicative skills in English as a Second Language for intending immigrants to Australia. They also concluded that G-theory and MFRA complemented one another: while G-theory provided general information to decide test design, and MFRA, on the other hand, provided specific information on individual examinees, raters, and items. These two early studies indicated the potentials of G-Theory and MFRA in the analysis of L2 performance assessment.

MFRA is adopted in several rating situations to investigate rater characteristics in L2 performance assessments. In Lumley and McNamara (1995), MFRA was adopted to investigate the stability of rater characteristics over a certain period. They set three rating occasions of the evaluations of a speaking test for health professionals. In the first two occasions, rater training was conducted to establish their reliability, and in the last occasion, no rater training was included. They made a comparison of rater characteristics among three occasions. The results showed the change of the rater characteristics through the three rating occasions, and Lumley and McNamara (ibid) concluded that the effect of rater training could not endure for long. This analysis was made possible by MFRA, estimating the severity of the raters independently of the data set.

Weigle (1998), furthermore, investigated the rater training effects in L2 essay writing. Sixty compositions in UCLA's English as a Second Language Placement Examination were evaluated by eight experienced and eight inexperienced raters with three evaluation items, rhetorical control, content, and language of 10-point scale. The rater training was conducted by the composition supervisor. In this rater training, the raters read "norming packets" with

sample compositions rated in the previous examination, compared their own rating with the rating in the previous examination, and lastly discussed the ratings with the supervisor. To investigate the effects of the rater training on the severity and the inconsistency of the experienced and inexperienced raters, the two sets of the evaluations were examined based on MFRA. Based on the comparison between the evaluations before and after the rater training, the following points were implied as the effects of the rater training. The raters tend to be in the similar levels of severity after the rater training, but this tendency was only for the inexperienced raters who showed extreme severity before the rater training. As for the experienced raters, almost no effect was found on rater variability in severity. The remarkable effect of the rater training, however, was the reduction of the inconsistency of raters' evaluation. This means that the individual raters evaluated the compositions more consistently after the rater training. Weigle (ibid) concluded that the rater training affects intra-rater reliability more strongly than inter-rater reliability.

MFRA is also applied in standard setting on performance assessment for certification in Japanese medical translation into English. In Kozaki (2004), the performances by trainees supervised by a translation expert were rated by translation experts and medical doctors. The raters evaluated the performance data along the analytic scales, such as schema conventions, information structure, grammar and vocabulary and graded pass-fail on the examinees. The ground rule of passing the examination was that at least three judges agreed to pass the examinee. Kozaki (ibid) firstly analyzed the evaluation ratings, based on the descriptive analysis and set the cut-off point of pass-fail in the analytic scale. The cut-off point in the analytic scales was the minimum of the average scores of the passers. In this analysis, some analytic scales were found to be against her assumption. In one analytic scale, the average score of the fail group was higher than that of the pass group. An examinee in the fail group obtained a higher score above the cut-off point than the others in the pass group.

Based on the information provided by MFRA, Kozaki (ibid) concluded that these results attributed to the inconsistency and the severity of the raters and the difficulty of the scales. This study is an example indicating the advantage of Rasch Analysis over CTT.

As the previous studies indicated, these two techniques are useful; while G-Theory detects relative effects of variability attributable to facets, MFRA provides information on specific elements of evaluation, raters, items and examinees. G-Theory allows investigators to handle sources of error in performance assessment, and it is possible that it predicts the dependability (reliability) according to conditions of manipulating the number of items and raters. In MFRA, examinees' ability is estimated independently from the severity of the particular raters and the difficulty of particular evaluation items. Examinees' ability is estimated in relation to the severity of raters and the difficulty of items. Moreover, the inconsistency of raters and items with the model can be excluded. In the present study, L2 performance evaluations were analyzed based on these two models. The analyses with G-Theory and MFRA were performed by the computer programs, GENOVA (Crick and Brennan, 1984) and FACETS (Linacre, 2006) respectively.

2.5 Reliability measurement

2.5.1 Reliability measurement in Classical Test Theory[1]

Reliability, which is generally examined by statistical analysis, is defined as the degree of coincidence of test scores when two or more tests are conducted to measure the same characteristic of examinees (Ikeda 1994). In this study, the reliability of performance evaluation by raters was examined by using G-Theory (Brennan, 1992). In this section, we review the reliability of measurement in CTT, and then, outline the concept and the procedure of G-Theory.

[1] This section is based on Ikeda (1994).

In CTT, it is assumed that a test score consists of true score and error as shown in the equation (2.1):

$$X = T + E \tag{2.1}$$

where X is an observed score; T, a true score; and E, an error of measurement. The correlation between the true score T and the error E is expected to be zero. As a result, the variance of the observed score is expressed as in the equation (2.2):

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{2.2}$$

where σ_X^2 is the variance of the observed score X; σ_T^2, that of the true score T; and σ_E^2, the error E. Under the hypothesis of the equation (2.1), the index of reliability is expressed as in the equation (2.3):

$$\rho_X = \frac{\sigma_T^2}{\sigma_X^2} \tag{2.3}$$

This index ρ_X is called coefficient of reliability. Since the variance of the true score is indeterminate, the various methods are adopted to estimate the reliability of measurement such as split-halves method, parallel test method, and so on. However, in CTT, the components of the error are not specified in a single analysis.

G-Theory, on the other hand, specifies multiple sources of measurement error in performance test. It is based on CTT and adopts the method of Analysis of Variance (ANOVA). Furthermore, the cost of performance evaluation can be estimated, such as number of raters and evaluation items.

2.5.2 Generalizability Theory[2]

G-Theory is a measurement model by which we can detect two or more sources of measurement error in test scores. In this section, the procedure of a two-crossed design in G-Theory is described. This design is of typical in performance assessment where we have two facets of measurements: raters and evaluation items. The analysis based on G-Theory consists of two steps: a generalizability study (G study) and a decision study (D study). The purpose of the G study is to estimate the relative effects of the respective sources of variance. In the D study, using the information of the variance components estimated in G study, we can assume the reliability of the test scores under several operational conditions. In this case, using the information estimated in the G study, we can assume the reliability of the test scores if we change the number of the items and the raters in the evaluation.

Suppose that examinees (e) do self-introduction task, and raters (r) evaluate the examinees performance using evaluation items (t). In the method of ANOVA, any observed score for a single evaluation item evaluated by a single rater can be expressed as:

$$X_{etr} = \mu + V_e + V_t + V_r + V_{et} + V_{er} + V_{tr} + V_{etr} \tag{2.4}$$

where μ is the grand mean in the population, and V stands for variances. Because of the orthogonality of each variance component, the population variance of X_{etr} can be deconstructed as:

$$\sigma^2(X_{etr}) = \sigma^2(e) + \sigma^2(t) + \sigma^2(r) + \sigma^2(et) + \sigma^2(er) + \sigma^2(tr) + \sigma^2(etr) \tag{2.5}$$

This is also represented in Figure 2.1 in terms of Venn diagram. Table 2.1 summarizes the

[2] This section is based on Brennan (1992), Ikeda (1994), and Bachman (2004).

expected mean squares and estimated variances of each variation factor.

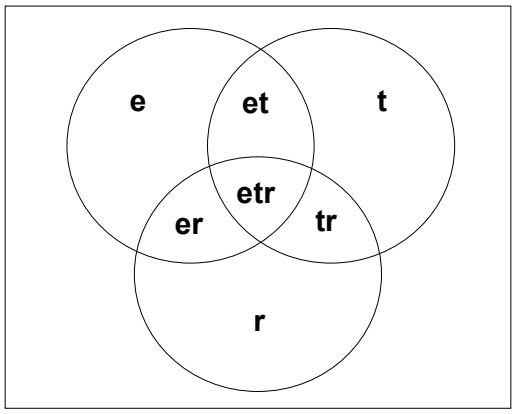

Figure 2.1 Venn Diagram for the Variances of Person, Task, and Rater Based on Brennan (1992)

Table 2.1 Expected mean squares and estimated variances based on Ikeda (1994)

Variation factor	Expected mean square (MS)	Estimated variance
e (examinee)	$\sigma^2_{etr} + n\sigma^2_{er} + r\sigma^2_{et} + nr\sigma^2_{e}$	$\hat{\sigma}^2_{e} = [MS_{e} - MS_{et} - MS_{er} + Ms_{etr}]/nr$
t (item)	$\sigma^2_{etr} + N\sigma^2_{tr} + r\sigma^2_{et} + Nr\sigma^2_{t}$	$\hat{\sigma}^2_{t} = [MS_{t} - MS_{et} - MS_{tr} + Ms_{etr}]/Nr$
r (rater)	$\sigma^2_{etr} + N\sigma^2_{tr} + n\sigma^2_{et} + Nn\sigma^2_{t}$	$\hat{\sigma}^2_{r} = [MS_{r} - MS_{er} - MS_{tr} + Ms_{etr}]/Nn$
Et	$\sigma^2_{etr} + r\sigma^2_{et}$	$\hat{\sigma}^2_{et} = [MS_{et} - MS_{etr}]/r$
er	$\sigma^2_{etr} + n\sigma^2_{er}$	$\hat{\sigma}^2_{er} = [MS_{er} - MS_{etr}]/n$
tr	$\sigma^2_{etr} + N\sigma^2_{tr}$	$\hat{\sigma}^2_{tr} = [MS_{tr} - MS_{etr}]/N$
etr (residual)	σ^2_{etr}	$\hat{\sigma}^2_{etr} = MS_{etr}$

Utilizing the information of the variance components specified in G study, the cost of performance evaluation can be estimated in D study. Adopting the model expressed in the

equation (2.4), the score of an examinee (e) is expressed as:

$$\mu_e = \mu + V_e = \tau_e \tag{2.6}$$

The difference between the grand mean and an observed score expressed in (2.7) is called absolute error:

$$\Delta_e = X_{etr} - \mu_e \tag{2.7}$$

Utilizing these variables, index of dependability (Φ) is defined as:

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\Delta)} \tag{2.8}$$

$\sigma^2(\tau)$ and $\sigma^2(\Delta)$ can be found by the variance components specified G study as:

$$\hat{\sigma}^2(\tau) = \hat{\sigma}_e^2 \tag{2.9}$$

$$\hat{\sigma}^2(\Delta) = \frac{\hat{\sigma}_t^2+\hat{\sigma}_{et}^2}{n'} + \frac{\hat{\sigma}_r^2+\hat{\sigma}_{er}^2}{r'} + \frac{\hat{\sigma}_{tr}^2+\hat{\sigma}_{etr}^2}{n'r'} \tag{2.10}$$

Utilizing the variance components specified in G study, Φ can be found by assigning value to n' and r' in the equation (2.10). This index is the G-Theory analogue of a reliability coefficient in CTT. This procedure helps us to estimate the cost of performance evaluation.

2.6 Item analysis

2.6.1 Item analysis in Classical Test Theory

In the examination of test items, we usually use two indices of item characteristics: correction rate and discrimination power. The correction rate of binary item p_j is defined as:

$$p_j = \frac{1}{N}\sum_{i=1}^{N} u_{ij} \tag{2.11}$$

where N is the number of examinees, and u_{ij} is examinees' responses: 0-1. The correction rate falls between 0 and 1, and an easy item obtains larger value. This rate is used as the index of item difficulty. The discrimination power is defined as the correlation coefficient between item responses and the sum of the test scores of examinees'. The items with high discrimination power are considered to reflect the sum of the test scores if the test is composed to measure a single trait. Although these two indices give useful information to test designers, there is fundamental limitation to the item analysis in CTT. This model depends on the abilities of the test takers and the test itself, even if the scores are normalized. The indices of item difficulty and item discrimination power totally depend on sampling population in CTT. Hence, we cannot predict the test results of a given learner in CTT. Transgressing the limitation, however, we can analyze the items based on IRT.

2.6.2 Item Response Theory[3]

IRT hypothesizes latent trait independent of examinee group. This trait is considered to be the same as factor one in factor analysis where items are dealt with as variables. In IRT, adopting cumulative normal distribution function, we can draw item characteristics curve (ICC) where y-axis indicates probability of correct response, and x-axis, the latent trait as

[3] This section is based on Toyoda (2002).

described the equation 2.12

$$\Phi\big(f(\theta)\big) = \int_{-\infty}^{f(\theta)} \phi(z)dz \tag{2.12}$$

To describe item difficulty, function of θ in 2.12 is defined as $f(\theta) = a(\theta - b_j)$, and ICC of a given item, $item_j$ is defined as $p_j(\theta) = \Phi\left(a\big(\theta - b_j\big)\right)$. This is called one parameter normal ogive model. In this equation, a is the constant value in one parameter model, and b_j is the item difficulty index. Because only b_j determines the property of the ICC, it is called one parameter model.

Since the equation 2.12 includes integral equation, the approximate formula (2.13) is used for convenience in which D is a scaling factor, 1.7. When D equals 1.7, it is noted that the discrepancy of estimated θ is below .01. This one-parameter logistic model is called Rasch model.

$$\int_{-\infty}^{f(\theta)} \phi(z)dz \cong \frac{1}{1+\exp\left(-Df(\theta)\right)} \tag{2.13}$$

In the actual situation, only available is examinees' responses, such as u'_i = [10110011]. In the estimation in IRT, fixing u_i, θ is estimated by optimizing the equation below.

$$L(u_i|\theta_i) = \prod_{j=1}^{n} p_j(\theta_i)^{u_{ij}}\, q_j(\theta_i)^{1-u_{ij}} \tag{2.14}$$

In this study, the evaluation scores are analyzed based on MFRA, which is an extension of Rasch model. It is adopted because item properties, trait level, and rater's severity can be separately estimated. The model is depicted in the equation below:

$$\log(P_{nmijk}/P_{nmijk}\text{-}1)=B_n - A_m - D_i - C_j - F_k \quad (2.15)$$

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

In MFRA in the present study, the Rating Scale Model was adopted, because the model assumes that the relative difficulties of the steps (intersections) within items (Embredson and Reise, 2000: 115). The model is expressed as follows (Embredson and Reise, ibid: 115-116):

$$P_x(\theta) = \frac{\exp[\psi_x + x(\theta - \lambda_i)]}{\sum_{x=0}^{M} \exp[\psi_x + x(\theta - \lambda_i)]} \quad (2.16)$$

where $\psi_x = -\sum_{j=0}^{x} \delta_j$ and $\psi_0 = \psi_m = 0$. δ_j is a category intersection parameter which describes each of the J = K – 1 category thresholds, and λ_i is a scale location parameter which expresses the relative difficulty of the particular item.

Raters and items can be excluded, based on the scores of infit calculated by MFRA. The score of infit "provides the size of the residuals, the differences between predicted and observed scores (McNamara, 1996: 172). The infit is the weighted mean-squared residual which is the index of unexpected responses near the point in which decisions are made. In

the case of raters, the infit of the raters indicates whether or not evaluations by the raters are inconsistent with the estimated ability of the examinees. The fit statistics produced by MFRA indicate the degree of individual raters' consistency in their ratings. An acceptable range of fit statistics can be fixed, but it depends on the context of the evaluation and the use of the results (Myford and Wolfe, 2004a and 2004b). The acceptable range of infit is "the mean ± twice the standard deviation of the mean score statistics" in the case where the population exceeded thirty (McNamara, ibid: 182). In this study, this criterion was adopted.

Kondo-Brown (2002) analyzed the assessment of Japanese L2 writing, based on MFRA. Three examinees out of 234 were identified as misfts (they obtained extremely high/low fit scores). She examined the examinees with high infit scores, and found out that two of them were children of Japanese immigrants: one was who had lived in Japan for several years, and the other was who demonstrated fluent and accurate expressions, but could write neither kana nor kanji and wrote the essay in alphabet. Kondo-Brown (ibid) eliminated these examinees in the subsequent analyses, because they were not candidates who the test developers had assumed as the examinees of the test. In MFRA, in this way, it is possible to detect an examinee based on the fit statistics.

2.7 Neural Test Theory[4]

NTT is a test theory which adopts the mechanism of self-organizing map (SOM: Kohonen, 2000). While interval scale is assumed as latent scale in IRT, ordinal scale is assumed in NTT. In this theory, examinees are grouped into some levels which a test developer sets up according to the probability estimated. Because NTT 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 NTT.

[4] This section is based on Shojima (2008).

The computational procedure of NTT is identical to that of SOM. Suppose that a test with nine items, and the number of the latent ranks set by the test analyzer is six. Figure 2.2 shows this example analysis of the test graphically. In this case, each latent rank has the reference vector with nine dimensions.

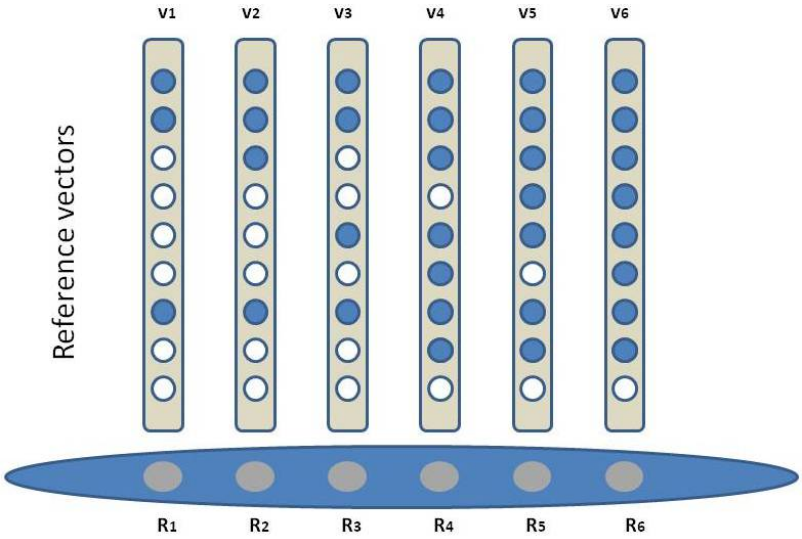

Figure 2.2 Image of the Computational Procedure of NTT

In NTT, 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 NTT is outlined below (Shojima, 2007):

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 neighboring nodes. (L5)

$V(t+1) \Leftarrow V_n^{(t)}$ (L6)

where T is learning time set by an analyzer; N, sample size; U, the response data of examinees, $U = \{u_i\}(i = 1,\ldots, N)$; 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):

$$R_w : w = \arg\min_{q \in Q} \| v_q^{(t)} - u_h^{(t)} \|^2 \quad (2.17)$$

where Q is the number of latent ranks set by the analyzer, and $v_q^{(t)}$ is the reference vector of a rank at the t-th period, $u_h^{(t)}$ is the response data of an examinee.

To examine the fit of data to the model in NTT, in addition to χ^2 statistics, several indices are available, but some of them provide with similar information for the degree of fit, and others provide with the information useful when two or more models are compared. In the present study, three indices are used to examine the degree of fit of data to the model: χ^2, the comparative fit index (CFI), and the root mean square error of approximation (RMSEA). These are indices for the fit of data which are not influenced by data size. It is generally accepted that above .90 of CFI and below .05 of RMSEA guarantees the goodness of fit of the data (Toyoda, 2007). The model fit statistics depend on the number of levels that test developers set up (Shojima, 2007). Hence, if we set up many levels such as thirty levels, the fit statistics indicates better fit of the data to the model compared to the case where we set up

small number of levels such as three. However, Shojima (2008) mentioned that because a test is not reliable enough to detect fine differences among the abilities of examinees', the realistic range of levels is from three to twenty. Shojima (ibid) also mentioned that levels should be set up not only by the model fit statistics, but also based on test developers' experience and the practicality of the test.

2.8 Studies on the relationship between proficiency and speech characteristics of L2 learners'

Attempts have been made in several fields of study to investigate the relationship between pronunciation and prosodic features of L2 learners' and their proficiency levels. In Second Language Acquisition (SLA), researchers focus on the influences of L2 experiences or L2 overall proficiency on learners' speech characteristics, and investigate learners' development of pronunciation and prosodic features, utilizing acoustic analyses of learners' speech. Typical of this sort of study in SLA are Trofimovich and Baker (2006 and 2007). They investigated the effect of L2 experiences on the development of suprasegmental features of Korean children, namely, fluency (e.g. speech rate, frequency and duration of pause) and prosody (e.g. stress timing and peak alignment). Another example is from Language Testing. In this field of research, the focus is on the differences in learners' speech characteristics according to task types in speaking assessment. Yuan and Ellis (2003) investigated the influences of task type of speaking test in learners' speech characteristics, namely fluency, complexity, and accuracy. Derwing, Rossiter, Munro, and Thomson (2004) and Foster and Skehan (1996) also examined the difference of learners' fluency according to task type in a speaking test. They adopted objective measures to examine learners' speech characteristics, such as the number and duration of pauses, and speech rate. Since the speech characteristics were manually measured in the field of SLA and language learning, extensive research has not been conducted.

Further investigations were done by speech technologists. The aim of their researches was to score the speaking performance of L2 learners' automatically by examining the relationship between human rating and speech characteristics. Neumayer, Franco, Digalakis, and Weintraub, (2000) collected the speech data of American learners of French and native speakers of French through a project. They also provided the speech data with human scoring, and investigated the strongest predictors of the human scoring among speech characteristics realized in the learners' speech. The focus of the rating is pronunciation. The highest scale indicates native-like pronunciation, and the lowest, strong foreign accent. In the correlation study, they used log-likelihood scores produced through Hidden Markov Model (HMM), two types of their normalized scores, and phone recognition errors in the speech recognizer as the quantitative measures of pronunciation. They found that the normalized scores of log-likelihood highly correlated with the human ratings, and implied the feasibility of automatic scoring of L2 pronunciation. In their results, furthermore, though their raters evaluated the learners' speech focusing on pronunciation, the index of speech rate was found to be the strongest predictor of the human rating. Cucchiarini, Strik, and Boves (2000a) investigated the relationship between fluency rating by experts on pronunciation and quantitative measures of fluency in read-aloud speech given by L2 learners and native speakers of Dutch. They collected the speech data through a read-aloud task through the telephone from sixty non-native speakers and twenty native speakers of Dutch, and their qualitative measures were calculated by using a continuous speech recognizer with trained acoustic models (39 Hidden Markov Models) with phonetically rich sentences read out by 4019 speakers. The fluency ratings on the speech data were made by not language teachers, but phoneticians and speech therapists, using a scale ranging from 1 to 10. The reliabilities of the ratings were examined from two aspects: inter-rater and intra-rater reliability, and both of them were found to be fairly high, though the raters were received no specific instruction

on the rating. The target qualitative measures were rate of speech, phonation/time ratio, articulation rate, number of silent pause, total duration of pauses, mean length of pauses, mean length of runs, number of filled pauses, and number of dysfluencies. These quantitative measures have been considered to be variables which affect fluency rating in the previous studies on L2 performance. In the correlation study, high correlations were found between the qualitative measures and the ratings: especially the correlation of speech rate with the rating was extremely high (.93). Cucchiarini, et al (2000a) assumed that the results were attributed to the type of speech they selected. In read-aloud speech, no variability in grammar and vocabulary is observed, but pronunciation still varies, which might have caused high reliability of the fluency ratings. The other differences in speech characteristics realized in the read-aloud speech between fluent speakers and non-fluent speakers of Dutch were the number of pause they make, rather than the length of individual pauses. Based on the results, Cucchiarini, et al (2000a) reported that speech rate was the best predictor of the fluency rating, because this measure was a complex variable of articulation rate and pause duration. In Cucchiarini, Strik, and Boves (2000b), they further investigated the relationship between the human ratings and the speech characteristics in read-aloud speech. They added a measure, log likelihood as the index of quality of segmental sounds. In the results of speech recognition, two types of information were obtained; one is starting and ending time of individual sounds, and the other is log-likelihood of individual sounds. Log likelihood is a measure indicating to what extent a segmental sound in question is similar to the counterpart in the acoustic model. If the model is trained with the speech given by native speakers of a language, and a target sound is produced by L2 learners of that language, the log likelihood in the results of speech recognition of the L2 learners' speech indicates the similarity between the native speakers' and the L2 learners' pronunciation of the target sound. Three characteristics in the read-aloud speech, total time duration, rate of speech, and log likelihood

were measured by an automatic speech recognizer, and their correlations with the human rating were calculated in three scales: segmental quality, fluency, and speech rate. In this study, again, rate of speech was found to be the best predictors of the human rating in all scales. Even with the segmental quality scale, the correlation of rate of speech was the highest among the three characteristics. Moreover, the index of pronunciation, log likelihood ratio was found to be a poor predictor of the human rating among the three characteristics, and the index of pronunciation and speech rate was closely related in their data. Based on the results, Cucchiarini et al. (2000b) concluded that in what we tried to measure in read-aloud speech, namely the construct of reading aloud, we could not separately measures these two characteristics, pronunciation and speech rate. In other words, fluent speakers tend to have good pronunciation. In addition to the investigation in read-aloud speech, Cucchiarini, Strik, and Boves (2002) made a comparison between read-aloud and spontaneous speech given by L2 learners of Dutch in the perspectives of fluency rating by human raters and quantitative measures of speech characteristics. The data for read-aloud speech were phonetically rich sentences produced by sixty L2 learners of Dutch, and the data for spontaneous speech were a part of a proficiency test for beginners and intermediate learners of Dutch, Profieltoets. This part of the test is a kind of discourse completion tasks. The learners were given a certain situation and indicated what they would say in that situation. The answers were relatively short. The two sets of data, read-aloud and spontaneous speech were credited by human raters in terms of fluency, and the quantitative measures were calculated: rate of speech, phonation/time ratio, articulation rate, number of sentence-internal pauses of no less than 0.2 second, total duration of pauses, mean length of pauses mean length of runs, number of filled pauses, and number of repetitions, restarts, and repairs. The results in this study showed that fluency was dependent on the speech type: learners tended to be less fluent in spontaneous speech than in read-aloud speech in terms of both the human rating and

the quantitative measures. Cucchiarini et al. (2002) pointed out that this difference was attributed to the requirement of cognitive load by tasks. In read-aloud speech, learners just read the text, but in spontaneous speech, learners need time to prepare for their answer to the question. This causes less fluent speech in spontaneous speech. Moreover, in spontaneous speech, articulation rate obtained almost no correlation with the human rating, though the other indices of speech rate were moderately correlated with the rating. On the other hand, both the articulation rate and the rate of speech in read-aloud speech were found to be good predictors of the fluency rating. The results, the authors assumed, were ascribed to the importance of pause in spontaneous speech. The articulation rate adopted in this study is the index of the average number of phonemes uttered in a certain period of time, which does not contain the information of pause, but the rate of speech, which adopted in this study, for example, was operationalized as the average number of phonemes divided by duration of speech including utterance internal silences. Cucchiarini et al. (2002) explained that in the speech where pauses become frequent, raters' attention to articulation would be reduced. Lastly, though it was not mentioned by the authors, the correlations of the quantitative measures with the fluency rating were relatively lower in spontaneous speech than those in read-aloud speech. Almost all the correlations in the read-aloud speech surpassed .80, but those in the spontaneous speech were around .40-60.

Through the investigations conducted by Neumayer et al. (2002) and Cucchiarini et al. (2000a, 2000b, and 2002), we can conclude that by using the speech characteristics automatically measured in read-aloud speech, we are able to construct a reliable automatic evaluation system of L2 read-aloud speech. In L2 read-aloud speech, the index of speech rate is a dominant predictor of human rating. The previous studies also indicate that raters' behavior were subject to task types (Neumayer et al., 2002; Cucchiarini et al, 2000a, 2000b, and 2002). In spontaneous speech, various characteristics can be found such as lexical

variety and syntactic structure, but these characteristics are fixed in read-aloud speech. Therefore, raters seemed to be hindered to concentrate on the characteristics related to fluency in the spontaneous speech.

Furthermore, the procedure and the evaluation of several in-service automatic scoring system of L2 speech were reported. Zechner, Higgins, Xi, and Williamson (2009) introduced an automatic scoring method. They used two sets of data from TOEFL Practice Online assessment (4162 responses) and TOEFL iBT Field Study (3502 responses). In the two sets of data, students were talking about everyday life and campus life. The speech data were scored by human raters in the range of 0-4. They selected several features in the speech, such as articulation rate, mean duration of pause, and pronunciation score calculated based on log-likelihood of segmental sounds in HMM. Then, they investigated the predictability of the human rating by the features, adopting two methods, multiple regression and classification and regression tree (CART). They compared the two methods and preferred multiple regression to CART, because multiple regression was simple and lucid, and moreover they found little difference in the results between the two methods. The correlations between the human raters and the machine scores were .57 for the data from TOEFL Practice Online assessment and .68 for the data from TOEFL iBT Field Study, compared to their inter-human rater agreement with the range of .74 to .94.

de Wet, Van der Walt, and Niesler (2009) investigated the predictability of human rating by speech characteristics in reading and repeating task using the data from L2 learners of English in a university in South Africa. The data of reading task were evaluated in scales of overall, degree of hesitation, pronunciation, and intonation with the range of 1-5, and the data of repetition task were evaluated in scales of overall, degree of success (the extent to which an examinee successfully imitate what he/she listen to) and accuracy with the range of 1-5. The speech characteristics they used were an index of goodness of pronunciation, which was a

variable transformed from log-likelihood of segmental sounds produced by HMM, rate of speech, and recognition accuracy by the speech recognizer. In the reading task, the rate of speech was found to be moderately correlated with the five scales (the correlation coefficients ranged from .46 to .53, while the correlation of goodness of pronunciation with the scales were fairly low, ranging from .02 to .13. In the repetition task, the relatively high correlations with the human ratings were found with the range of .59-.71. Bernstein, De Jong, Pisoni, and Twonshend (2000) introduced an automatic scoring of spoken English, PhonePass SET-10, which consisted of five parts: reading, repeating, saying opposites, giving short answers to some questions, and giving open answers to some questions. They demonstrated the good correlations of the scores in PhonePass with other well-known English test, such as TOEFL and TSE (.73 and .88, respectively). Furthermore, they showed the good correlation of the PhonePass scores with the six levels in CEFR. PhonePass scores from 2.0 to 3.9 predict level "below level A1," from 4.0 to 4.9 predict level A1, from 5.0 to 5.5 predict level A2, from 5.6 to 6.1 predict level B1, from 6.2 to 6.7 predict level B2, from 6.8 to 7.2 predict level C1, and from 7.3 to 8.0 predict level C2. These results also indicate the possibility of the predictability of human ratings by the speech characteristics realized in read-aloud speech.

However, in these studies, less attention has been paid to raters' characteristics, rating procedure, and examination of scores based on test theory, which can be considered to affect the scoring. In the automatic scoring methods depicted above, the scores credited by human raters were predicted by speech characteristics. Hence, the reliability of human rating must be examined before the correlation studies between the scores and the speech characteristics. Furthermore, they have not mentioned about the scoring and feedback method in an automatic evaluation system to be built or in service. To predict evaluation scores credited by human raters by using speech characteristics in speech, we need to calculate the correlation between

them. In the previous studies the rating scores were dealt with as interval scale to obtain the correlations. However, as the predicted scores, it is questionable that interval scale is useful for examinees. Test scores at least have to work as benchmark by which the examinees clearly understand their own ability and what to do to proceed to the next stage in the target ability. Without can-do descriptions, test scores in interval scale are difficult for examinees to understand. Therefore, from an educational point of view, test scores in ordinal scale with can-do descriptions should be produced even in the automatic scoring for the sake of clear understanding of test score by examinees. As Eskenazi (2009) mentioned, this field of research benefits from the knowledge in computer science, statistics and signal processing as well as in second language acquisition, cognitive science and linguistics, and yet requires the knowledge in language testing especially in the examination of reliability in human rating and the method of scoring and feedback.

In the L2 performance evaluations conducted in the present study, the applicability of the criterion for the evaluations, namely CEFR, is investigated to the context of English language learning in Japan, and the reliability and the consistency of raters and evaluation items are examined, based on the test theories, G-Theory and MFRA. In the evaluation of read-aloud speech, furthermore, a text with a certain length is adopted for raters to catch the learners' speech characteristics. The raters in the present study are L2 users, namely, Japanese teachers of English with experience, considering the present situation of English language learning. The purpose of the present study is to investigate the correlation between the scores delivered through this evaluation and the speech characteristics, to construct an automatic evaluation system based on the results of the correlation studies, and to examine the reliability of the score predicted by the system.

3 Applicability of Common European Framework of References in the context of Japan[5]

3.1 Introduction

As discussed in 2.2, wordings in the "Can-do" statements in European Language Portfolio (ELP) should be contextualized to measure learners' language proficiency precisely, and the levels in Common European Framework of References (CEFR) are selected according to the context of language learning. In Hasselgreen (2005), to use the "Can-do" statements in the assessment of young language learners in Nordic/Baltic countries, expressions in the "Can-do" statements were modified and the levels in CEFR were selected, because the statements were written and the levels in CEFR were set for adolescent and adult learners of language. Accordingly, the "Can-do" statements in ELP and the levels in CEFR need to be examined in order to apply them in other context of language learning. This chapter reports the study which investigated the applicability of the levels in CEFR and the "Can-do" statements in ELP in the context of English language learning in Japan.

3.2 Method

3.2.1 Participants

Participants were students and teachers in a course, which is called "Tutorial English" at Waseda University. In this course, each class consists of four or five students and one teacher. The students are encouraged to discuss in English from their daily life to more complex matters according to their levels. The levels are "Beginners", "Basic", "Pre-Intermediate", "Intermediate", "Pre-Advanced", and "Advanced", which refer to the six

[5] This chapter first appeared in Tsutsui, E., Kondo, Y., & Nakano, M. (2007) An investigation on criterion for assessment of speaking ability of Japanese learners of English with reference to Common European Framework of References [Nihonjin Eigo Gakushusha No Jissenteki Hatsuwa Noryoku Ni Kansuru Hyoko Kijun No Kento Common European Framework of References O Kiban Tosite]. *Proceedings of the 5th Annual Conference of the Japan Association for Research on Testing*. 88-91.

levels in CEFR. The textbooks and the tasks in the classrooms were composed with reference to CEFR. The students are placed on these six levels according to the results of a placement test, Web-based Test for English Communication (WeTEC), which was created for the placement of students in this course. WeTEC is made up of four sections: vocabulary, idiomatic expressions, listening, and dictation to measure communication ability in English.

The teachers were native speakers of English or near native speakers of English. Among the students who attended this course, 2619 students self-evaluated their speaking ability, and their teachers in charge evaluated their 982 students among them using the "Can-do" statements selected from Schärer (2004). Table 3.1 shows the number and the levels of students in this study.

Table 3.1 The number and the levels of students

Levels	Number of students
Beginners	32 (13)
Basic	417 (153)
Pre-Intermediate	591 (225)
Intermediate	601 (229)
Pre-Advanced	704 (266)
Advanced	274 (96)
Sum	2619 (982)

Note. The numbers in brackets indicate the number of students who received the evaluation by their teachers.

3.2.2 The "Can-do" statements in European Language Portfolio

Ninety nine "Can-do" statements were selected from Schärer (2004). Because the study focused on the speaking ability of learners', the statements were selected from four categories: spoken interaction, spoken production, language strategy, and language quality. In this study, students self-evaluated their speaking ability, and their teachers evaluated the ability of their students' with these ninety nine "Can-do" statements. The "Can-do" statements for the students were translated into Japanese by the author and the co-authors[6]. The students used the translated version of them, and the teacher used the original version with 4-point scale: "Can hardly do it", "Cannot really do it", "Can do it to some extent", and "Can do it". Since the students have been already categorized into six levels based on the results of the placement test, four types of questionnaires were composed according to the levels. A student in a level self-evaluated their speaking ability with the "Can-do" statements of their own level and of the two adjacent levels: the upper and the lower levels. For example, Questionnaire B includes the "Can-do" statements in the levels of A2, B1, and B2.

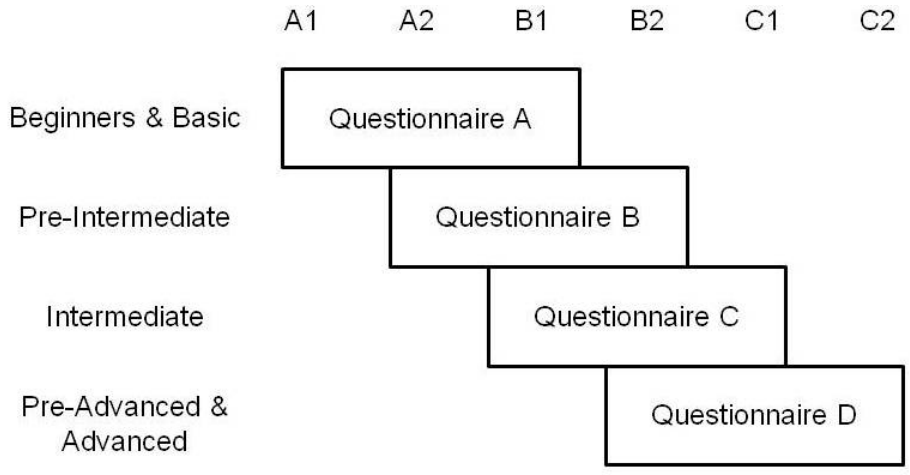

Figure 3.1 Construction of Questionnaires

[6] Kazuharu Owada, Associate Professor at Ritsumeikan University, Eiichiro Tsutsui, Associate Professor at Hiroshima International University, and Michiko Nakano, Professor at Waseda University

A questionnaire contains common “Can-do” statements with the adjacent questionnaires. The construction is depicted in Figure 3.1. The original and translated versions of the “Can-do statement” are listed in Appendix A. The evaluations were done with printed questionnaires on the last day of the courses. The teachers evaluated their students with the same questionnaire as their students used.

3.2.3 Analysis

The self-evaluation by the students and the evaluation by the teachers were analyzed based on Classical Test Theory (CTT) and Item Response Theory (IRT), two-parameter logistic model. The responses to the “Can-do” statements were in 4-point scale, but in the analyses, the four category are converted to binary data: “Can hardly do it” and “Cannot really do it” to 0, and “Can do it to some extent” and “Can do it” to 1. In the analysis based on IRT, all the items in the four questionnaires were equated by using the common items. The analysis of IRT was performed by the computer software, BILOG-MG 3 (Zimowski, Muraki, Mislevy, and Bock, 2003). The correct response ratio and the point biserial correlation coefficients of items were calculated based on CTT, and the difficulties and the discrimination powers of items were calculated based on IRT. To investigate the applicability of the “Can-do” statements in ELP and the levels in CEFR, the means of the item difficulty and the discrimination power in each level in CEFR were calculated, and the differences were examined in item difficulty between the evaluation by the students and the teachers.

3.3 Results

The correct response ratio and the point biserial correlation coefficients, the difficulty and the discrimination power based on IRT were calculated. They are listed in Appendix B. Table 3.2 and 3.3 shows the means of the difficulty and the discrimination power of the “Can-do”

statements in each category in the self-evaluation by the students and the evaluation by the teachers. The item characteristic curves drawn in Figures 3.2 and 3.3 are the average item characteristic curves in each level. The six levels were clearly differentiated in both the self-evaluation and the evaluation by the teachers; no curve in the graphs is crossed. As for the discrimination power, the values in the evaluation by teachers were higher than the self-evaluation by the students, but the mean of the item difficulty in the evaluation by teachers were statistically higher than the self-evaluation by the students ($t(173) = 2.08$, $p = .05$ (two-tailed)).

Table 3.2 The means of the difficulty and the discrimination power of the "Can-do" statements in each category in the self-evaluation by the students

	A1	A2	B1	B2	C1	C2
Difficulty	-1.11	-0.89	-0.04	0.53	1.13	1.26
Discrimination power	0.78	0.81	0.90	0.98	1.30	1.40

Table 3.3 The means of the difficulty and the discrimination power of the "Can-do" statements in each category in the evaluation by the teachers

	A1	A2	B1	B2	C1	C2
Difficulty	-1.11	-0.93	-0.29	0.14	0.65	0.81
Discrimination power	2.05	1.80	1.75	1.97	2.17	2.22

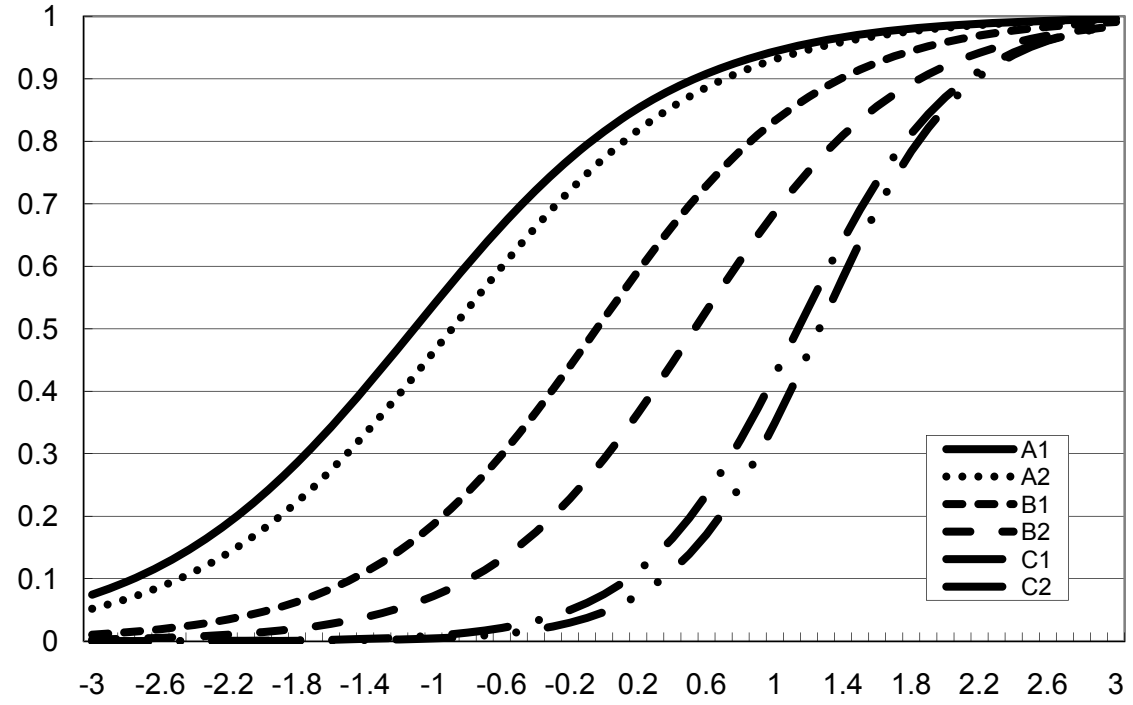

Figure 3.2 Average Item Characteristic Curves in Each Category in the Self-evaluation by Students

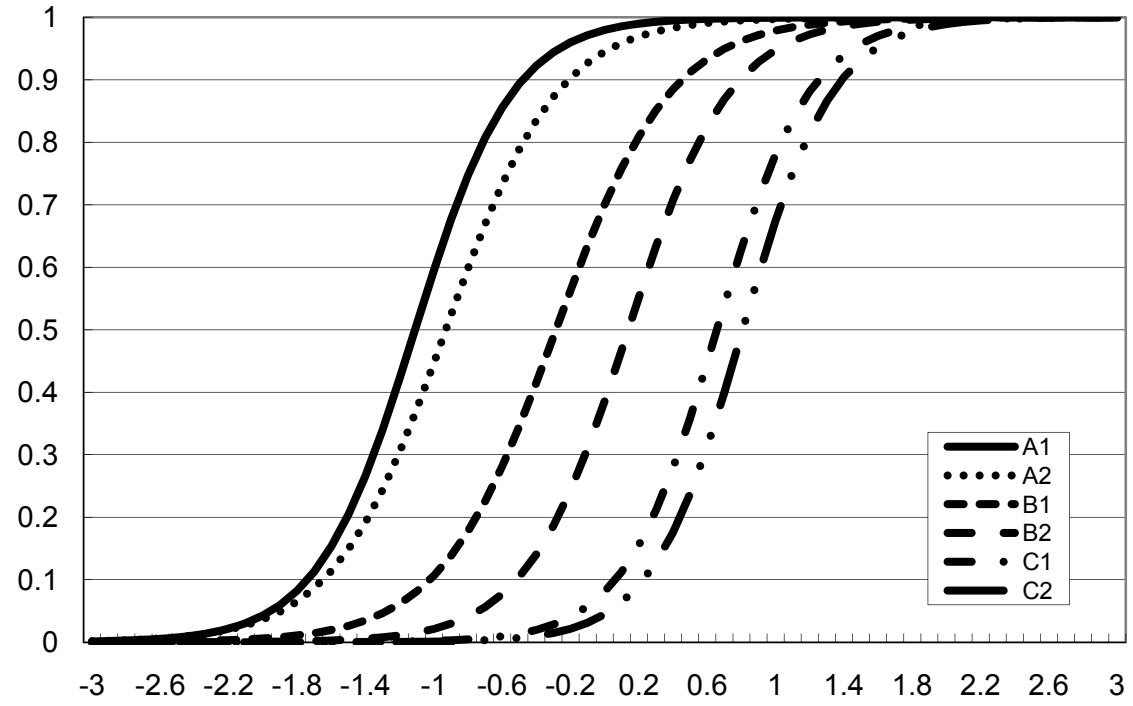

Figure 3.3 Average Item Characteristic Curves in Each Category in the Evaluation by Teachers

The correlation coefficient of the item difficulty between the self-evaluation and the evaluation by the teacher was fairly high (r = .85), but the differences of the item difficulty was found in some items between the evaluation by the teachers and the self-evaluation by the students. The mean of the absolute values of the differences was 0.48, and the standard deviation was 0.41. Therefore, the items with the difference more than the twice standard deviation from the mean were defined as the items with big difference between the evaluation by the teachers and the self-evaluation by the students in this analysis. Five items listed below were found to be items with the big difference. In these items, the students were harsher than the teachers. In other words, the students thought that the tasks in the "Can-do" statements below were much more difficult than the teachers thought.

1. I can make simple transactions in shops, post offices or banks.
2. I can maintain a conversation or discussion but may sometimes be difficult to follow when trying to say exactly what I would like to do.
3. I can say when I don't understand.
4. I can very simply ask somebody to speak more slowly.
5. I have sufficient vocabulary to express myself on matters connected to my field and on most general topics.

The "Can-do" statements above can be categorized into three: problems related to conversation requiring its schema information, the cultural differences between European countries and Japan, and the translation. As for 1, to make transaction in post offices and banks, usually schema information in those contexts is required. For example, the students might have thought that there were several differences between English speaking countries and Japan when they open a savings account in a bank. Furthermore, they might not know

the differences between a savings account and a checking account. They might not know the vocabulary used in this context. Generally speaking, a few Japanese students have experience of staying in English speaking countries. These might have made the students think that the task was very difficult. However, the teachers might have thought it was not a difficult task for their students because they talked about general matters freely in their classroom. The second problem realized in Items 3 and 4 is the cultural differences. The task described in these two items was requesting speech act. The students must know the expressions to be used when they do not understand what their conversational partners say, such as “Can you repeat that again please?” or “Did you say…?”, but in some context, requesting is thought to be impolite in Japanese culture, especially to elderly people. Requesting is expected to be done polite, and the use of language is sometimes difficult for the students even in Japanese. That might have caused the students to think the task was very difficult. The last problem was related to the translation. In item 2, the word, discussion was translated into “giron” in Japanese, but “discussion” in this item is used in the phrase “a conversation or discussion”. This means that “discussion” indicates “talking over some issues”, “a casual discussion”, but “giron” in Japanese rarely used in such a context. The students might have thought that “giron” (discussion) in this item was more serious one. The same was found in Item 5. “my field” in this item was translated into “senmon” which means specialty in Japanese. The students might have thought that it was fairly difficult to discuss some issues related to their specialty. The teachers, on the other hand, might have focused on the expression “express myself on matters connected to my field”. The teachers might have thought that it was not so difficult to express on matters connected to their fields.

3.4 Summary and discussion

The aim of this study was to investigate the applicability of the levels in CEFR and the "Can-do" statements in ELP in the context of language learning in Japan. 2619 Japanese university students self-evaluated their speaking ability, and their teachers evaluated their 982 students among the 2619 students with the "Can-do" statements in ELP. Their evaluation was analyzed based on IRT, two parameter logistic model. The results indicate, as shown in Figures 3.2 and 3.3, that the six levels in CEFR were clearly differentiated in both the self-evaluation and the evaluation by the teachers; no curve in the graphs is crossed. The high correlation of the item difficulty between the evaluation by the teachers and the self-evaluation by the students also implies the applicability of the levels in CEFR and the "Can-do" statements in ELP in the context of language learning at Waseda University. It indicates that native and near native speakers of English and Japanese language learners of English share the degree of difficulty of the task which the "Can-do" statements describe. However, discrepancy was found in the difficulty of some of the "Can-do" statements between the students and the teachers. The source of the discrepancy might lie in the problems not related to the speaking ability of the students. It was interpreted that linguistic problems related to schema information that a task described in a "Can-do" statement requires the cultural differences between Japan and European countries, and the translation. The discrepancy in the item difficulty between the teachers and the students, however, were found in only five items. We can reduce the discrepancy by re-translating the "Can-do" statements and by improving our language teaching materials and method where students can learn schema information in some contexts and the cultural differences between Japan and English speaking countries. Based on the discussion, we can conclude that the levels in CEFR and the "Can-do" statements in ELP are applicable to the context of language learning at Waseda University where English is learnt as a foreign language.

4 Rater training effect in L2 performance evaluation[7]

4.1 Introduction

As discussed in 2.4, the two approaches to L2 performance evaluation, Generalizability Theory (G-Theory) and Multifaceted Rasch Analysis (MFRA) have been used as complementary methods to investigate the reliability of the evaluation and the consistency of raters' evaluation. The previous studies indicated the usefulness of the information on the evaluation produced by these two methods in the reliability examination in L2 performance evaluation (Lumley and McNamara, 1995).

In the evaluation reported here, raters evaluated recorded self-introduction speech made by Asian learners of English before and after rater training. The purpose of the study is to investigate the effect of rater training. The study focuses on the change of reliability of the evaluation applying the information provided by G-Theory and the changes of raters' internal consistency, and also investigates the change of raters' consistency and severity, applying the information on raters' behaviors produced by MFRA through the rater training.

4.2 Method

4.2.1 Participants

Seventy three Asian learners of English participated as an examinee in this study. Their first languages are Thai, Japanese, Korean, Tagalog, Mandarin, and Taiwanese. They are graduate or undergraduate students. Their L2 background is summarized in Table 4.1.

Five Japanese raters with the master degree of Applied Linguistics participated in this study. Their average year of learning English was 18.3 with S.D. 6.5 and that of teaching English,

[7] A part of this chapter first appeared in 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*. 8 pages in CD-ROM Proceedings.

10.9 with S.D. 8.7. Their experiences of teaching English were not only in primary, secondary, and high school and university in Japan, but also some of them have taught English as an L2 to non-Japanese learners. Language teachers of non-native speakers of English were chosen, because of their knowledge on the context of learning English (See 2.3).

Table 4.1 Key information of the participants in self-introduction task

	M	*SD*	Range
Age	20.77	3.14	13
Study of English (year)	10.38	3.94	22

N = 73.

4.2.2 Recording procedure

All the recording was made in soundproof rooms in the universities which the participants belonged to. The participants 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, the participants gave their self-introduction to an interviewer, and the interviewer only gave approving nods. 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.

4.2.3 Rating procedure

Evaluation items were selected from those in Yashiro, Araki, Higuchi, Yamamoto, and Komissarov (2001), and each item was thoroughly reviewed in order to make the items suitable in the evaluation of unprepared L2 speech. The items are depicted in Table 4.2.

Table 4.2 Evaluation items in self-introduction

1. Loudness	9. Speech rate	17. Paralinguistic cues
2. Sound pitch	10. Prosody	18. Confidence
3. Quality of vowels	11. Fluency	19. Try to sound cheerful
4. Quality of consonants	12. Place of fillers	20. Try to sound friendly
5. Epenthesis	13. Frequency of fillers	21. Grammatical accuracy
6. Elision	14. Place of pause	22. Coherency
7. Word stress	15. Frequency of pause	23. Absence of tension
8. Sentence stress	16. Length of silent pause	24. Foreign accentedness

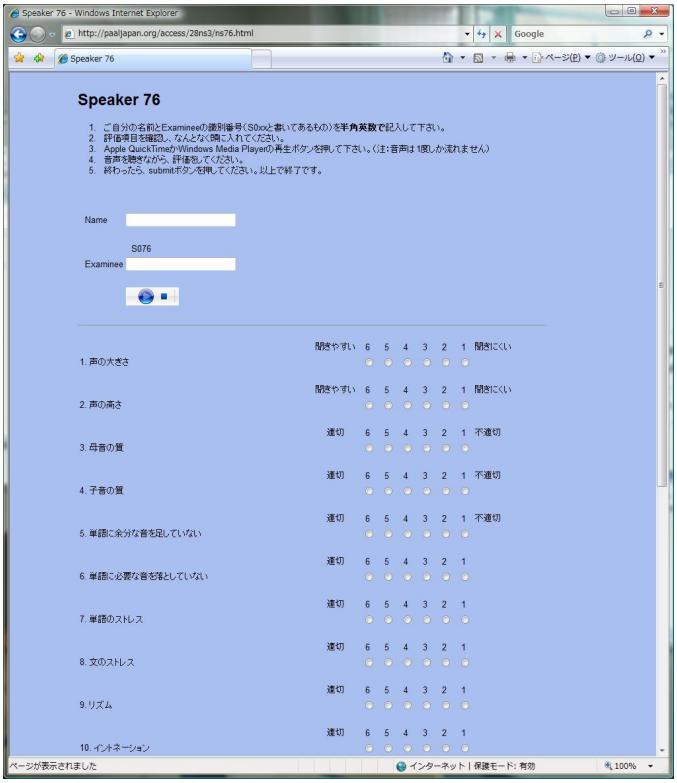

Figure 4.1 A Sample of the Evaluation Website

The raters evaluated the participants' speech on the website individually. On the website, the raters listened to and evaluated the recorded participants' speeches in view of overall

proficiency and twenty four subcategories of overall proficiency where a 6-point Likert scale was adopted. A sample of the evaluation website is shown in Figure 4.1. All the raters evaluated every speech in this evaluation in the same order.

4.2.4 Rater training procedure

Rater training was conducted according to the manual provided by Council of Europe (Council of Europe, 2003). The procedure of linking a test to CEFR consists of five steps: Familiarization, Specification, Standardization training and benchmarking, Standard setting, and Validation (Council of Europe, ibid: 10-11) in the present study. Firstly, raters received the overview of the speech data. They are unprepared self-introduction speech and recorded in the universities where the speakers belonged to, and the speakers were Asian learners of English who were graduate or undergraduate students. Then, the raters discussed the speech characteristics of the learners' and selected the evaluation items from Yashiro, et al (2001), listening to a couple of speech data. This stage is "Specification" of the evaluation in the manual. After the discussion on the speech characteristics of the learners, the raters were given the descriptors and the levels in CEFR and watched the video (North and Hughes, 2003) which depicted the learners divided into six levels. This stage is "Familiarization" to the descriptors and the levels in CEFR. Lastly, the raters discussed the descriptors and the levels in CEFR, watching the video, and discussed the characteristics of learner language in each level. This is the stage of "Standardization Training and Benchmarking" and a part of "Standard Setting" in the manual. Rater training was conducted three times during two weeks. This activity led the raters to establish the images of the learners of six levels.

4.3 Examination of rater training effects based on Generalizability study

In this section the effect of rater training are reported in terms of reliability improvement. In

the present study, the raters and the items were a random facet, because they could be exchanged with other raters, and evaluation items were also exchangeable, which could be taken from any other items related to the L2 speech assessment. All the examinees were evaluated by all the raters. Hence, the design of the Generalizability study (G study) was examinee × items × raters. The design of the G study is a random effect model with two facets: twenty four items and five raters, which assumes that the raters and the items interacted interchangeably. The focus of this study is dependability (reliability) of test scores with full facets. The estimated variances of each facet (e.g. examinee, item, and rater) were examined, and the indices of dependability were compared before and after our rater training. In this experiment, fifteen learners randomly selected from the participants described in 4.2.1 participated as examinees, and five language teachers described in 4.2.1, as raters.

Tables 4.3 and 4.4 show the results of the G study before and after the rater training. Comparing the estimated variances before and after the training, the examinees' ability accounts for 43 per cent and 63 per cent, and the rater related variables, for 12 per cent and 8 per cent. A remarkable difference before and after our rater training is the difference 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 these after the training. In the rater training our raters watched the video where the learners of six levels were depicted. It must have helped the raters to clarify how they should scale.

Table 4.3 G study before the rater training

	SS	*df*	*MS*	EV
e (examinee)	1518.55	14	108.47	0.43
t (item)	1886.89	23	82.04	0.51
r (rater)	425.41	9	47.27	0.12
et	1056.77	322	3.28	0.28
er	296.39	126	2.35	0.08
tr	651.91	207	3.15	0.18
etr (residual)	1419.09	2898	0.49	0.49
Sum	7255.01	3599	247.05	2.08

Note: *SS*: sum of squared deviation, *df*: degree of freedom, *MS*: Mean square, EV: Estimated variance.

Table 4.4 G study after the rater training

	SS	*df*	*MS*	EV
e (examinee)	2221.45	14	158.68	0.63
t (item)	342.15	23	14.88	0.08
r (rater)	440.75	9	48.97	0.11
et	397.76	322	1.24	0.08
er	814.28	126	6.46	0.25
tr	351.33	207	1.70	0.09
etr (residual)	1177.84	2898	0.41	0.41
Sum	5745.66	3599	232.32	1.66

Note: *SS*: sum of squared deviation, *df*: degree of freedom, *MS*: Mean square, EV: Estimated variance.

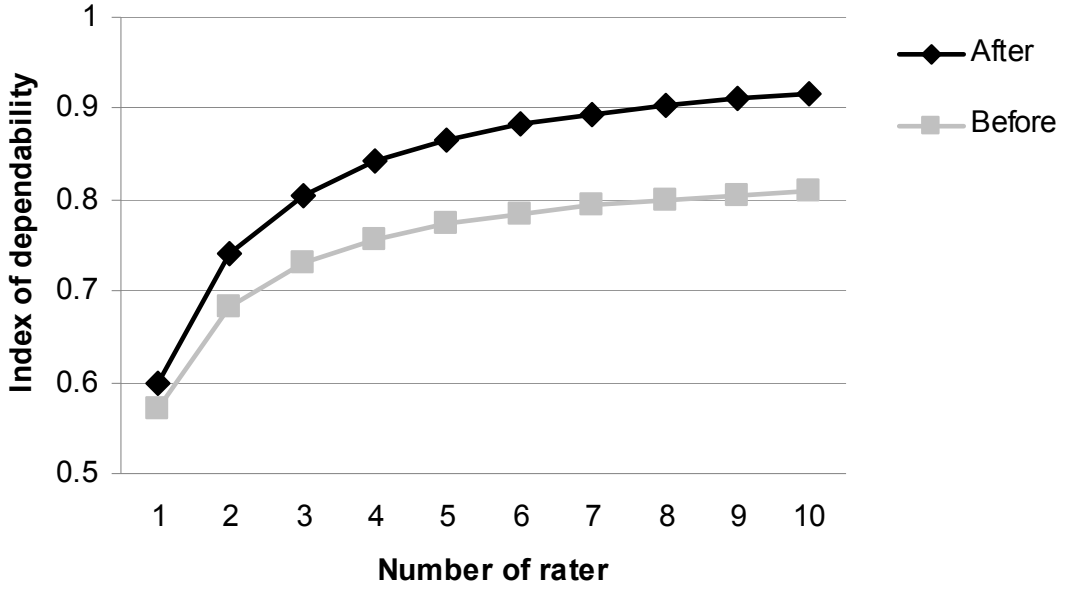

Figure 4.2 Change of Index of Dependability

Utilizing the information of the estimated variance specified in the G study, the indices of dependability Φ were calculated before and after the training. Generalizability coefficient is an index only for examinee-related factors, and the index of dependability, on the other hand, is an index for all variation factors including examinee-related factors. The former is larger than the latter. In the results of the D-studies, according to conditions; manipulating the number of items and raters, we can predict dependability in several conditions. By comparing the results of the D-studies before and after the rater training, the cost reduced by the rater training is revealed. Evaluation conditions were simulated where one to ten rater(s) evaluated examinees using one to ten evaluation item(s). All the simulations are found in Tables 4.5 and 4.6. The evaluation condition of one to ten rater(s) using the ten items is described in Figure 4.2. With the acceptance that this index is the analogue of a reliability coefficient, the minimum value is .85 for a reliable evaluation. The change of Φ described in Figure 4.2 is the simulation where one to ten rater(s) evaluate(s) examinees using ten items. If the rater training is conducted, above 0.85 of Φ can be obtained by only four raters.

Table 4.5 The index of dependability before the rater training

		Rater									
		1	2	3	4	5	6	7	8	9	10
Item	1	.21	.27	.29	.31	.31	.32	.33	.33	.33	.34
	2	.33	.40	.44	.46	.47	.48	.48	.49	.49	.50
	3	.40	.49	.53	.55	.56	.57	.58	.58	.59	.59
	4	.45	.54	.58	.61	.62	.63	.64	.65	.65	.65
	5	.48	.58	.63	.65	.67	.68	.68	.69	.70	.70
	6	.51	.61	.66	.68	.70	.71	.72	.72	.73	.73
	7	.53	.64	.68	.71	.72	.73	.74	.75	.75	.76
	8	.55	.65	.70	.73	.74	.75	.76	.77	.77	.78
	9	.56	.67	.72	.74	.76	.77	.78	.79	.79	.80
	10	.57	.68	.73	.76	.77	.78	.79	.80	.80	.81

Table 4.6 The index of dependability after the rater training

		Rater									
		1	2	3	4	5	6	7	8	9	10
Item	1	.39	.52	.59	.63	.66	.68	.69	.70	.71	.72
	2	.48	.62	.69	.73	.76	.78	.79	.80	.81	.82
	3	.52	.67	.74	.77	.80	.82	.83	.84	.85	.86
	4	.55	.69	.76	.80	.82	.84	.85	.86	.87	.88
	5	.56	.71	.77	.81	.84	.85	.87	.88	.88	.89
	6	.57	.72	.78	.82	.85	.86	.88	.89	.89	.90
	7	.58	.73	.79	.83	.85	.87	.88	.89	.90	.91
	8	.59	.73	.80	.83	.86	.88	.89	.90	.90	.91
	9	.59	.74	.80	.84	.86	.88	.89	.90	.91	.91
	10	.60	.74	.81	.84	.87	.88	.89	.90	.91	.92

4.4 Examination of rater training effects based on MFRA

In this experiment, fifteen learners were randomly selected from the participants described in 4.2.1, and the evaluations given by five raters were analyzed. The evaluation scores before and after the rater training were independently analyzed based on MFRA. In the process of the analysis of the evaluation scores before the rater training, three items, “Paralinguistic cues”, “Absence of tension”, and “Foreign accentedness”, were found to be extremely inconsistent evaluations items, whose infits surpassed 3.00. Hence, these three items were excluded in this analysis. Table 4.7 shows the infits and the severity measures of the raters and the infits and the difficulty measures of the evaluation items before and after the rater training respectively.

Table 4.7 Infits and severity of raters before and after rater training

	Before training		After training	
	Infit	Severity	Infit	Severity
Rater 1	1.14	-0.74	1.24	-1.33
Rater 2	1.11	-0.16	1.22	0.49
Rater 3	0.95	-0.40	0.91	-0.48
Rater 4	0.93	0.12	0.86	-0.09
Rater 5	0.73	-0.47	0.82	-0.07
M	0.97	-0.33	1.01	-0.42
SD	0.16	0.32	0.20	0.67

The logit values of the severity before and after the rater training need to be adjusted to make them comparable with each other (Lumley and McNamara, 1995). Adding -0.09 to the each value of the severity before the training, the two sets of the severity were compared by t-test. There were no difference in the severity measure of raters before and after the training ($t(4) = 0.56$, $p = .60$ (two-tailed)). As for the index of the self-consistency in the raters, the infit, no inconsistent raters were found both before and after the training in the condition that the upper and lower limit of the fit statistics are set to 1.4 and 0.6, respectively (Wright and Linacre, 1994).

4.5 Summary and discussion

The purpose of the study was to investigate the effects of rater training in an L2 performance evaluation. Rater training was conducted in order for raters to clearly understand the criteria, the evaluation items, and the evaluation procedure. In the training, the raters watched the videos (North and Hughes, 2003), and discussed the learners' characteristics at each level.

The analyses of the evaluations were done before and after the rater training based on G-Theory and MFRA. In the analyses based on G-Theory, the variance related to the items was reduced to about one sixth after the training, though no difference was found in the rater characteristics before and after the training in the analysis based on MFRA.

These results might be mainly attributed to the background of the raters in this study. The raters in these evaluations are familiar with the context of learning English in Asia. They also know the learners themselves. It is the reason why the raters were self-consistent before the training. In the analysis by Weigle (1998), inexperienced raters tended to be self-inconsistent, while experienced raters were self-consistent before the training. However, as the results of G study in the present study shows, the variance related to the evaluation items were reduced after the training. This is because the raters might have understood the contents of the evaluation items better through the training. In performance evaluation, the difficulty and the consistency of the evaluation items are greatly influenced by raters' understanding of the contents of items. In the present study, no difference were found in raters' characteristics in the results of MRFA, but the variance related to the evaluation items were found to be reduced in the results of G study. This can be said to be one of the effects of the rater training.

The other finding of this study is about the eligibility of the raters whose first language is not English in L2 performance evaluation. Comparing the results of Kim (2009) with those of the present study, our raters were equally self-consistent with the raters of native speakers of English in Kim (2009). Furthermore, it is legitimate to adopt L2 users as the raters, because, in countries where English is a foreign or second language, the non-native users teach and learn English. In this situation, teachers of L2 users are the most appropriate in L2 performance evaluation if they are self-consistent in their ratings.

The raters in this study were Japanese language teachers of English, though the learners'

speech data were collected widely from Asia. If raters share their first language with learners, it may influence on their evaluation. The answer to this question could not be found in the results of the present study.

5 Investigation of objective measures as predictors in self-introduction speech[8]

5.1 Introduction

As discussed in 2.3, the rater training is one of the important processes in L2 performance assessment. Although the results in Lumley and McNamara (1995) indicate that the sustainability of rater training is weak, Weigle (1998) points out that raters show similar levels of severity when they are extremely severe or lenient in their evaluation and become more consistent in their evaluations after rater training. In the present study, the raters who received the rater training described in Chapter 4 evaluated the spontaneous speech of seventy-three Asian learners of English. Their evaluations were analyzed to detect unreliable raters and items using Multifaceted Rasch Analysis (MFRA) to obtain reliable scores in the evaluation. Thereafter, the relationship between the scores and speech characteristics in the spontaneous speech was examined to discover score predictors among speech characteristics.

The previous studies, summarized in 2.4, have found that if the number of variables is increased, the correlations between the scores given by human raters and the speech characteristics decreases. In the repetition task, learners mimick the pronunciation of the target utterance so that speech characteristics such as intonation and rhythm will be kept constant. In the reading task, because no pronunciation model is presented, intonation and rhythm can vary, depending on individual learners, but syntactic features and vocabulary are kept constant. In spontaneous speech, learners' utterances cannot be controlled for all

[8] The early version of this chapter appeared in Tsutsui, E., Kondo, Y., Owada, K., Ueda, N., & Nakano, M. (2006, October). *Exploring communication abilities of English language learners in the eastern Asian context: From the perspectives of Common European Framework and World Englishes*. Paper presented at The 12th Annual Conference of the International Association for World Englishes, Nagoya, Japan and Nakano, M., Kondo, Y., Tsutsui, E., & Owada, K. (2007, June). Daigaku eigo kyouiku ni okeru koutou happyou nouryoku no hyouka to sokutei. [Evaluation and Measurement of Second Language Speech in the University Context: Towards an Automatic Evaluation System]. Paper presented at 2007 Convention of the Japan Association of College English Teachers Kanto Chapter, Tokyo, Japan.

aspects of language use.

In the previous studies from which these insights were drawn, however, the raters were users of the target language; for example, in Neumayer, Franco, Digalakis, and Weintraub (2000), native speakers of French evaluated the speech of American learners of French, and in Cucchiarini, Strik, and Boves (2000a, 2000b, and 2002), Dutch phoneticians and speech therapists evaluated the speech of learners of Dutch. Considering the context of learning English in Asia, native English speakers are not necessarily eligible to evaluate the speech of Asian learners of English. In Asia, L2 users of English teach English to L2 learners. Therefore, L2 users (Japanese English language teachers) joined as raters in our evaluation. It is natural for L2 English learners in Asia to be evaluated by teachers who use English as a second language (See 2.3).

The purposes of this study are to obtain reliable scores in the evaluation of spontaneous speech spoken by L2 English users and to examine rater behaviors in the evaluation of spontaneous speech by investigating the relationship between evaluation scores and speech characteristics. The analysis of the evaluation based on MFRA is first described, and the correlation analysis is then reported.

5.2 Method

5.2.1 Participants

The participants were seventy-three Asian learners of English, and the raters were ten Japanese English language teachers. All the raters received the rater training described in Chapter 4. The language background and the information on English study of the participants are described in 4.2.1. In this evaluation, the ten raters scored the speech of all seventy-three learners in the same order. In the correlation analysis, thirty learners were randomly selected, and their speech was objectively analyzed.

5.2.2 Recording and evaluation procedure

Recording and evaluation were performed using the same procedure described in Chapter 4, sections 4.2.2 and 4.2.3. Recording took place in sound proof rooms located at the universities to which the learners belonged. Recorded speech was evaluated by the raters individually through a website that had been set up for the evaluation. Raters used twenty-four evaluation items described in 4.2.3 in the evaluation of learners' spontaneous speech.

5.2.3 Analysis

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" (McNamara, 1996: 172). The acceptable range of infit is "the mean ± twice the standard deviation of the mean score statistics" in cases when the population exceeds thirty (ibid: 182). The analyses below were repeated to meet this standard.

5.3 Rater and item selection based on Multifaceted Rasch Analysis

The data were analyzed three times. In the first analysis, two items, Item 23 "Absence of tension" and Item 24 "Foreign accentedness," exceeded the acceptable range ($0.99 \pm 0.28 \times 2 = 0.43 - 1.55$), and no rater exceeded the acceptable range ($1.01 \pm 0.38 \times 2 = 0.25 - 1.77$). These two items were excluded and the data were reanalyzed. In the second analysis, two items, Item 17 "Paralinguistic cues" and Item 18 "Confidence," exceeded the acceptable range ($0.99 \pm 0.17 \times 2 = 0.64 - 1.34$), and no rater exceeded the acceptable range ($1.05 \pm 0.38 \times 2 = 0.24 - 1.79$). These two items were excluded, and the data were reanalyzed again. In the third analysis, no item and no rater exceeded the acceptable range ($1.00 \pm 0.36 = 0.28 -$

1.72 and 0.98 ± 0.13 = 0.73 – 1.24, respectively). Although the range of raters' severity is wide, the infit scores of the raters fell into the acceptable range. The results indicate that the severity of the raters' and the difficulty of the items vary, but they were in the acceptable range of self-consistent in the model of MFRA. The same can be said for the item property. The details of rater and item measurement reports are depicted in Appendix C. Table 5.1 shows the rater measurement report and Table 5.2 shows the item measurement report of the last analysis. These reports list the severity of the raters, the difficulty of the items, model error, and infit. The second column indicates the severity of the raters. In the third column, the standard error is presented. The fourth column shows the infit scores. Figure 5.1 summarizes this analysis, plotting the severity of the raters, the ability of the examinees, and the difficulty of the items in a scale. The range of the severity of the raters is relatively wide, while that of the difficulty of the items is narrow.

Table 5.1 Rater measurement report in the self-introduction task

Rater	Severity	Error	Infit
Rater 1	0.65	0.03	1.14
Rater 2	-0.97	0.03	0.87
Rater 3	0.24	0.03	0.49
Rater 4	-0.22	0.03	0.65
Rater 5	-0.74	0.03	1.07
Rater 6	-0.38	0.03	1.43
Rater 7	-0.34	0.03	1.02
Rater 8	0.02	0.03	0.51
Rater 9	-0.62	0.03	1.22
Rater 10	-1.99	0.03	1.64
Mean	-0.44	0.03	1.00
SD	0.69	0.00	0.36

Table 5.2 Item measurement report of the self-introduction task

Item	Difficulty	Error	Infit
1. Loudness	-0.80	0.04	1.15
2. Sound pitch	-0.84	0.04	1.00
3. Quality of vowels	0.17	0.04	0.87
4. Quality of consonants	0.19	0.04	0.91
5. Epenthesis	0.16	0.04	1.10
6. Elision	0.10	0.04	1.02
7. Word stress	-0.15	0.04	0.74
8. Sentence stress	0.01	0.04	0.76
9. Speech rate	0.19	0.04	0.96
10. Prosody	-0.14	0.04	0.89
11. Fluency	0.18	0.04	1.14
12. Place of fillers	0.22	0.04	0.98
13. Frequency of fillers	0.44	0.04	1.09
14. Place of silent pause	0.14	0.04	1.04
15. Frequency of silent pause	0.28	0.04	1.10
16. Length of silent pause	0.18	0.04	1.13
19. Try to sound cheerful	-0.13	0.04	0.95
20. Try to sound friendly	-0.24	0.04	0.83
21. Grammatical Accuracy	-0.03	0.04	0.95
22. Coherency	0.07	0.04	1.18
Mean	0.00	0.04	0.99
SD	0.32	0.00	0.13

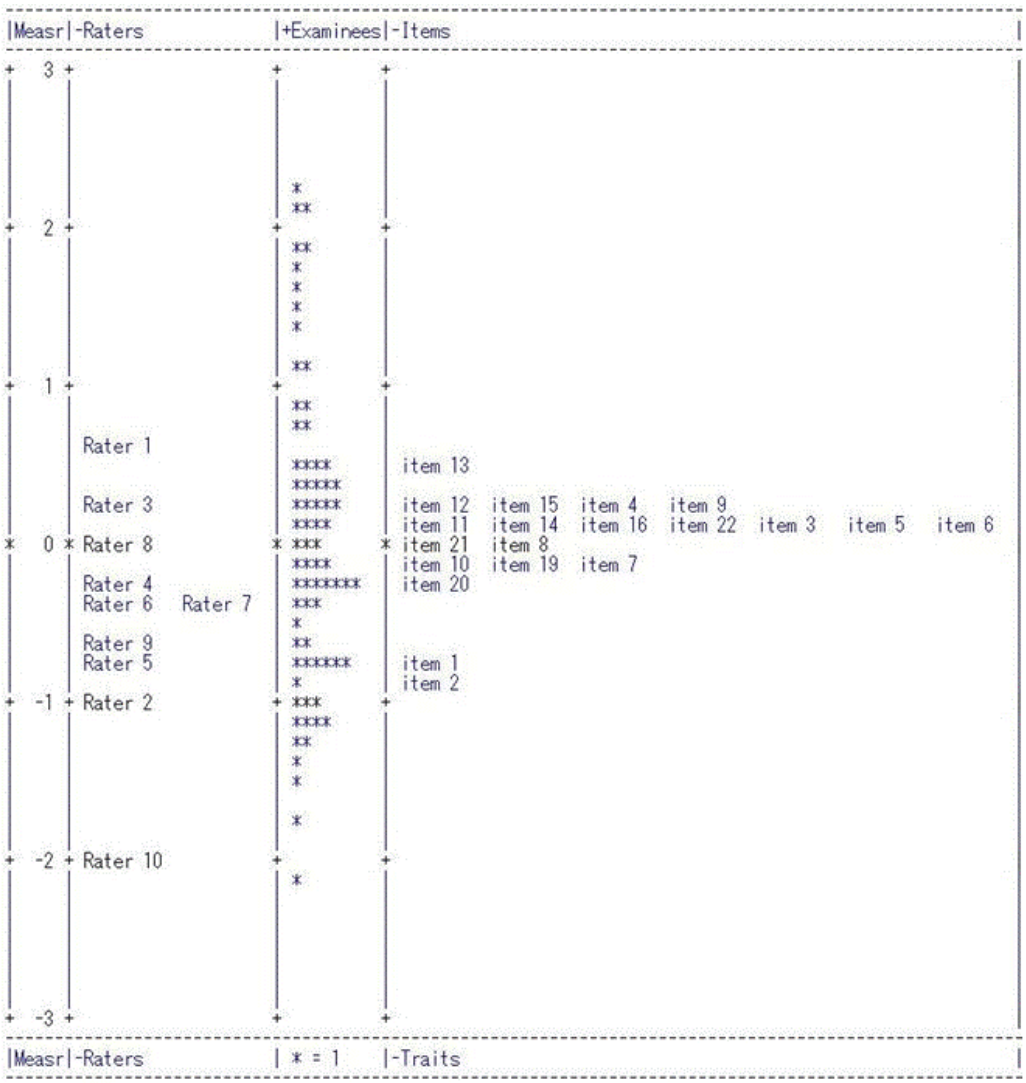

Figure 5.1 Vertical Yardstick by FACETS in Self-introduction Task

5.4 The predictability of the objective measures in the self-introduction

The objective measures were scrutinized by using multiple regression analysis (stepwise method). The criterion variable is the evaluation score, and the predictor variables are the objective variables selected from the objective measures of L2 speech characteristics adopted in the previous studies on the speech characteristics of L2 learners: Ano (2001); Foster and Skehan (1996); Iwashita, McNamara, and Elder (2001); and Yuan and Ellis (2003). The measures were defined as follows. The transcription and the coding of the speech data was

done by the authors and the co-authors[9].

Number of silent pauses: The number of silent pauses a learner made in his or her speech. Silence that is 250 ms or longer is considered to be a silent pause.

Number of fillers: The number of fillers a learner made in his or her speech. "mm," "erm," and "uh" are considered to be fillers. These are extracted based on subjective judgment.

Total length of silent pauses: The total amount of time that silent pauses and fillers account for in the speech.

Average length of fillers: The mean length of fillers during an amount of time.

Average length of silent pauses: The mean length of silent pauses during an amount of time.

Frequency of silent pause and fillers: Calculated by the formula Y/(1+X), where Y is the total number of words, and X is the total number of silent pauses and fillers.

Words per minute: The average number of words spoken in one minute.

Type/token ratio: The percentage of word types in a speech

Ratio of easy words and proper words: The percentage of easy words in a speech; "easy words" are defined based on JACET 8000 (Daigaku Eigo Kyoiku Gakkai Kihon Go Kaicho Iin kai, 2003).

Ratio of error-free C-unit: The learners' errors were analyzed based on Owada (2005).

Words per C-unit: The average number of words in a C-unit.

Flesch Reading Ease: This index indicates the readability of a passage. The score is calculated by the formula 206.84 – 1.015 × (average sentence length) – 84.6 × (average number of syllables per word).

[9] Kazuharu Owada, Associate Professor at Ritsumeikan University, Eiichiro Tsutsui, Associate Professor at Hiroshima International University, and Norifumi Ueda, Lecturer at Mejiro University.

The significance level of the F value of the predictor variables is set to five percent, and variables above this level are excluded. In summary, the equation below expresses the relationship between the evaluation score and the objective measures.

$$\hat{Y} = X_1 + X_2 + \ldots + X_n \quad (5.1)$$

where $\hat{Y}$ is the evaluation score, and $X_1 + X_2 + \ldots + X_n$ are objective measures such as words per minute, number of fillers, and the ratio of error-free C-units.

Table 5.3 The correlation coefficients between all the variables

	a	b	c	d	e	F	g	h	i	j	k	l	m	n
a	1	.29	-.22	-.60	-.26	-.68	.75	.79	-.01	-.36	.18	-.05	.40	.21
b		1	.05	-.04	-.31	-.53	-.07	.26	-.11	-.42	.28	-.39	.38	.00
c			1	-.24	-.39	-.01	-.28	.09	-.25	-.13	.10	-.01	-.07	-.04
d				1	.25	.81	-.64	-.82	.20	.36	-.28	-.10	-.24	-.03
e					1	.33	-.04	-.36	.37	.22	-.04	-.13	-.13	-.03
f						1	-.55	-.80	.16	.45	-.36	.16	-.40	.01
g							1	.83	-.20	-.39	.24	-.03	.49	.18
h								1	-.28	-.62	.38	-.05	.51	.19
I									1	.13	-.01	-.13	-.26	.10
j										1	-.63	.14	-.41	-.20
k											1	-.19	-.10	.09
l												1	-.36	.10
m													1	-.06
n														1

Note.

a: evaluation score

b: number of silent pause

h: word per minute

i: type toke ratio

c: number of filled pause

d: sum of silence length

e: average length of filled pause

f: average length of silent pause

g: frequency of pause

j: ratio of easy words

k: ratio of proper nouns

l: ratio of error-free C-unit

m: words per C-unit

n: Flesch Reading Ease

Table 5.3 shows the correlation coefficients between all the variables. In the multiple regression analysis, two predictor variables were found to be significant: words per minute and number of filled pause. The significance of the model was verified ($F_{(2, 27)} = 34.21$, $p < .01$). These two variables could predict the criterion variable to a large extent (adjusted $R^2 = .70$). The details of these two variables are shown in Table 5.4. Table 5.5 shows the correlation coefficients between these predictor variables and the criterion variables.

Table 5.4 Predictor variables of the evaluation score

Predictor variable	*β*	*p*
Word per minute	.82	.01
Number of Filled pause	-.29	.01

Table 5.5 Correlation coefficients between the objective measures and the evaluation

	1	2	3
1. Evaluation score	1	.79	-.22
2. Word per minute		1	.09
3. Number of Filled pause			1

According to the correlation coefficients between the two predictor variables and the criterion variable, it is possible to interpret the influence of these two predictor variables to

the criterion variables separately (Toyoda, 1998: 44-45). In this case an approximate equation can be obtained:

$$R^2 \simeq \alpha_1 + \alpha_2 \tag{5.2}$$

where α_1 is the partial correlation coefficients between one predictor variable, and α_2, that of the other predictor variable. Assigning the values in this study to this equation, we have the equality below:

$$0.7 \simeq 0.82^2 + 0.29^2 = 0.67 + 0.08. \tag{5.3}$$

As described above, the evaluation score can be predicted by using the words per minute variable at a rate of about sixty-seven percent, and the number of filled pauses variable at a rate of about eight percent. About twenty-five percent of the evaluation score is influenced by unknown factors. Because the correlation coefficient between words per minute and the number of filled pauses is not zero, this analysis is not a strict examination of the partial correlation coefficients.

5.5 Summary and discussion

In this study, an evaluation of spontaneous speech by trained raters was analyzed based on MFRA. Four items were found to be inconsistent. The four excluded items were "Paralinguistic cues," "Confidence," "Absence of tension," and "Foreign accentedness." As for the first item, since the data included few samples of paralinguistic cues such as coughing and laughing it may have been difficult for the raters to use this item. As for the second and third items, since the speech data that the raters evaluated were not visual data, the raters

could not judge the speech data in terms of these two points. The last item, "Foreign accentedness" does not seem to be related to proficiency level. Unlike a language that is used by the great majority of native speakers and the minority of the L2 learners, such as Japanese, the number of the L2 learners exceeds that of the native speakers of the English language (Crystal, 2003). English is now used by a large population of L2 speakers. Therefore, "Foreign accentedness" might not be directly related to the proficiency of L2 speakers, and as such cannot be used in the evaluation of L2 speech. After the four inconsistent evaluation items were excluded, no rater surpassed the standard for the acceptable range of consistency. This result is attributed to the rater training and the raters' knowledge of the context of learning English in Asia. The experienced raters who received rater training and the analysis of the evaluation based on MFRA were the factors that contributed to the reliable scores in this evaluation.

The second finding of this study was that the evaluation scores could be predicted to a large extent by two indices of learners' speech characteristics. As Table 4.4 shows, the correlation coefficients of timing-control characteristics such as the indices of pause control and speech rate were higher than other variables that expressed the vocabulary size and syntactic complexity and accuracy (e.g., ratio of easy words, type-token ratio, Flesch Reading Ease, and Ratio of error-free C-units). The dominant factors in the evaluation of spontaneous speech are timing-control characteristics, not syntactic accuracy and complexity or lexical variety. We can assume that the speech data of L2 learners are suitable for evaluating timing-control characteristics, but the data are not suitable for evaluating syntactic features or vocabulary richness. This study found out substantial predictors of evaluation scores given by human raters. These two speech characteristics are measured by computer. By using these two characteristics, we can predict the evaluation score given by human raters. Automatic speech evaluation system is another possibility of speech evaluation for L2

learners.

In addition, as shown in Table 5.3, moderate correlations with the evaluation scores were found in an index of lexical richness, “ratio of easy words,” and an index of syntactic complexity, “words per C-unit.” Generally, raters evaluate learners’ timing-control characteristics, syntactic complexity, and lexical richness at the same time when evaluating spontaneous speech. This complex work is difficult for raters to complete, even if they receive rater training. This might be the cause of the low correlations between speech characteristics and the human ratings reported by the previous studies described in 2.4.

6 Investigation of objective measures as predictors in read-aloud speech

6.1 Introduction

The results of the evaluation of the spontaneous speech reported in Chapter 5 indicated the possibility of automatic second language (L2) speech evaluation system, because the statistically significant predictors of the evaluation by human raters are the speech characteristics which can be measured by computer. Although the study reported in Chapter 5 did not focus on phonetic features such as pronunciation, other than timing control characteristics, it is assumed that the human ratings are related to the phonetic features of learners. Hence, 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-aloud speech. First, the evaluation of L2 read-aloud speech is examined based on Multifaced Rasch Analysis (MFRA). The five pilot studies are then reported. The speech data used in the analyses were taken from Asian English speech database described in Chapter 8.

6.2 Method

6.2.1 Participants

Each participant 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 6.1 shows the key information of the participants. Five raters joined this evaluation; they were Japanese language teachers who had participated in the rater training described in Chapter 4, and their reliability had been examined in the evaluation of spontaneous speech that was reported in Chapter 5. The raters

evaluated all the speeches that were read by the 101 Asian English learners, and in the five pilot studies, read-aloud speeches were randomly selected and used to investigate the relationship between the evaluation scores and the speech characteristics.

Table 6.1 Key information of the participants in read-aloud speech

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

Note. $N = 101$.

6.2.2 Recording and evaluation procedure

Recording and evaluation were performed using the same procedure as described in Chapters 4 and 5: see details in 4.2.2 and 4.2.3. Recordings were made in sound proof rooms located in the universities to which the learners belonged. Before the participants read a text out loud, they read the text silently to understand the context of the text.

Table 6.2 Evaluation items in read-aloud speech

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

They were also given the time to ask an interviewer questions about the contents and the vocabulary of the text. Recorded speeches were evaluated by the raters individually through

a website that had been established for the evaluation. To evaluate the read-aloud speech, the fourteen evaluation items in Table 6.2 were selected from the evaluation of the spontaneous speech.

6.2.3 Text

The text that the participants 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(NIE) corpus (Deterding and Ling, 2005), and is used in the phonetic description of the International Phonetic Association. Below is the passage used in this study.

> 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.

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 Jones, 2003).

6.3 Rater and item selection based on Multifaceted Rasch Analysis

The analysis was completed using the same procedure as was used in the evaluation of spontaneous speech in Chapter 5. Raters and items were excluded based on their infit score. In the present analysis, neither raters nor items exceeded the acceptable range (the mean ± twice the standard deviation of the mean score statistics): see Appendix D. The rater measurement report is shown in Table 6.3 and the item measurement report is shown in Table 6.4.

Table 6.3 Rater measurement report in read-aloud speech

Rater	Severity	Error	Infit
Rater 1	-0.82	0.03	1.30
Rater 2	-0.40	0.03	0.84
Rater 3	-0.87	0.03	1.00
Rater 4	-0.17	0.03	0.96
Rater 5	0.31	0.03	0.68
Rater 6	-0.90	0.03	1.21
Mean	-0.48	0.03	1.00
SD	0.44	0.00	0.21

Figure 6.1 summarizes this analysis, plotting the severity of the raters, the ability of the examinees, and the difficulty of the items in a scale. A comparison of these results with those of the spontaneous speech showed that the range of the severity of the raters was narrow: the raters' severities were much the same as they had been in the evaluation of spontaneous speech. The mean of the severity for the spontaneous speech was –0.44, and that of the present study was –0.48. As for the fit statistics, the range was narrower in the present results than for those of spontaneous speech (0.84–1.30 and 0.49–1.6, respectively).

The same tendency was found in the results of the item analysis. The means of the item difficulty were identical (0.00), and the range of the fit statistics was narrower in the evaluation of the read-aloud speech than in the spontaneous speech (0.74–1.18 and 0.88–1.17, respectively): see 5.4.

Table 6.4 Item measurement report of read-aloud speech

Item	Difficulty	Error	Infit
1. Loudness	-0.65	0.05	1.00
2. Sound pitch	-0.76	0.05	0.95
3. Quality of vowels	0.38	0.04	0.97
4. Quality of consonants	0.21	0.04	0.94
5. Epenthesis	0.05	0.04	1.14
6. Elision	0.23	0.04	1.17
7. Word stress	-0.15	0.04	1.15
8. Sentence stress	0.27	0.04	0.88
9. Rhythm	0.47	0.04	1.01
10. Intonation	0.51	0.04	0.91
11. Speech rate	-0.14	0.04	0.95
12. Fluency	-0.14	0.04	0.95
13. Place of pauses	-0.14	0.04	0.95
14. Frequency of pauses	-0.14	0.04	0.95
Mean	0.00	0.04	0.99
SD	0.37	0.00	0.09

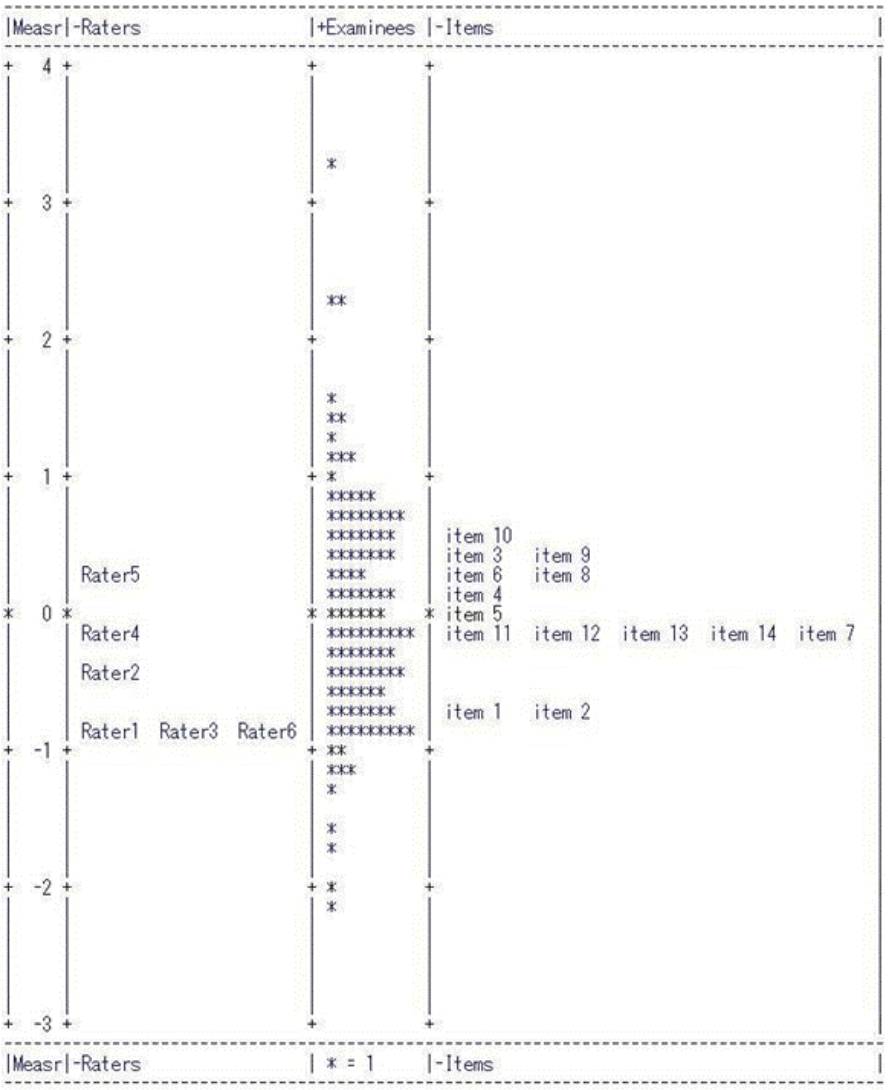

Figure 6.1 Vertical Yardstick by FACETS in read-aloud speech

6.4 The predictability of the objective measures in read-aloud speech

6.4.1 Speech timing-control characteristics[10]

The relationship between the evaluation scores and speech timing control characteristics was examined in the read-aloud speech given by eight Japanese, eight Korean, seven Chinese, three Thais, three Filipinos, and two Malaysians. There were thirty-one participants, and their average age was 23.19 years old with a SD of 3.63. Their average time studying

[10] This section first appeared in 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.

English was 13.48 years with a SD of 4.84. The speech timing control characteristics analyzed were the number and the duration of nonlexical and silent pauses, mean length of run, the number of unneeded syllables, pruned syllables per second, and the average ratio of weak syllables to strong syllables. These objective measures were selected from Munro and Thomson (2004); Trofimovich and Baker (2006 and 2007); Riggenbach (1991); and Towell, Hawkins, and Bazergui (1996).

Pauses are divided into two categories: nonlexical pauses and silent pauses. A nonlexical pause is defined as a nonlexical word that is not found in the written text that the participants read aloud. Silence was counted every 10 ms from 10 ms to 400 ms, and the correlations between these silences and the evaluation scores were examined. For example, silences above 10 ms were counted, and the correlation was examined between these silences in the first step; in the next step, silences above 20 ms were counted and the correlation was examined between these silences. Figure 6.2 shows the correlation coefficients between these silences and the evaluations scores. The correlation of the silences with the evaluation scores is the highest when a silent pause is defined as a silence above 100 ms. Therefore, a silent pause is defined as a silence beyond 100 ms.

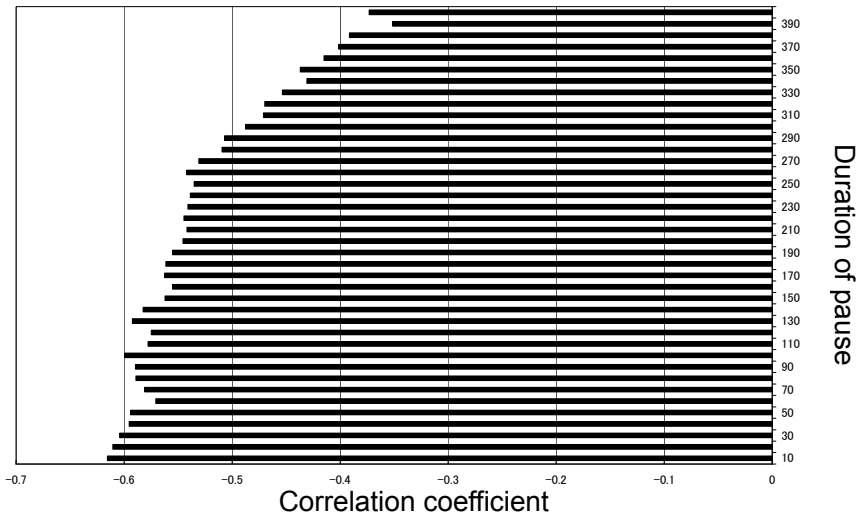

Figure 6.2 Correlation between the Silent Pauses and the Evaluation Score

The "mean length of run" is defined as the average number of syllables that learners uttered between two silent pauses or two nonlexical pauses. As for the speech rate indices, three indices were examined: syllables per second, articulation rate, and pruned syllables per second. Syllables per second is the total number of syllables, including self-correction, self-repetition, and filled pauses, divided by the total number of seconds. Articulation rate is the total number of syllables, including self-correction, self-repetition, and filled pauses, divided by the total number of seconds, excluding silent pauses. Pruned syllables per second is the total number of syllables, excluding self-correction, self-repetition, and filled pauses, divided by the total number of seconds. This index is based on Munro and Thomson (2004). Judging from the correlation coefficients of the three speech rate indices with the evaluation score, pruned syllables per second was adopted as the index of speech rate. Pruned syllables per second are operationalized as follows:

$$S = (T - E) / TD \tag{6.1}$$

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 (Riggenbach, 1991). The ratio of unaccented syllables to accented syllables is operationalized as follows:

$$R = A / U \tag{6.2}$$

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).

Table 6.5 shows the mean, standard deviation, and correlation coefficient of the speech characteristics with the evaluation scores.

Table 6.5 Evaluation score and the mean, standard deviation, and correlation coefficient of the speech characteristics.

	M	*SD*	*r*
Number of nonlexical pauses	0.22	0.50	-.38
Number of silent pauses	19.41	8.23	-.62
Duration of nonlexical pauses	0.03	0.08	-.23
Duration of silent pauses	9.34	4.09	-.56
Mean length of run	7.87	2.25	.67
Number of unnecessary syllable	2.58	2.68	-.51
Pruned syllables per second	3.24	0.55	.74
The ratio of weak syllables to strong syllables	0.52	0.09	-.43

Table 6.6 shows the correlation coefficients between the speech characteristics and the evaluation score. Multiple regression analysis was performed; the predictor variables were the speech characteristics and the criterion variable was the evaluation score. Two predictor variables were found to be statistically significant: Pruned syllables per second and the ratio of weak syllables to strong syllables. The significance of the model was verified ($F_{(2, 28)} = 31.59$, $p < .01$). The model indicates that the evaluation score can be predicted by these two variables to a large extent (adjusted $R^2 = .67$). The equation below was obtained:

$$\hat{Y} = 1.92a_1 - 6.57a_2 - 2.79, \quad (6.3)$$

where a_1 is the pruned syllables per second and a_2 is the ratio of weak syllables to strong syllables. As shown in Tables 6.5 and 6.6, pruned syllables per second resulted in the highest correlation coefficient with the evaluation score, and the ratio of weak syllables to strong syllables resulted in a moderate correlation coefficient with the score; however, the latter index independently correlates with the scores from the other variables. The correlation coefficients of the ratio of weak syllables to strong syllables were fairly low with the other variables than the evaluation scores.

Table 6.6 The correlation coefficients between the speech characteristics and the evaluation score

	ES	NFP	NSP	DFP	DSP	MLR	SPS	RWS
ES	1	-.38	-.62	-.23	-.56	.67	.74	-.43
NFP		1	.35	.86	.37	-.31	-.49	.08
NSP			1	.15	.90	.90	-.76	.21
DFP				1	.18	-.18	-.36	.05
DSP					1	-.75	-.82	.15
MLR						1	.78	.19
SPS							1	.05
RWS								1

Note.

ES: evaluation score

NFP: number of filled pauses

NSP: number of silent pauses

DSP: duration of silent pauses

MLR: mean length of run

SPS: syllable per second

DFP: duration of filled pauses RWS: ratio of weak syllables to strong ones

These results led us to conclude that the human rating in the evaluation of read-aloud speech can be predicted to large extent by the two previously mentioned speech characteristics. This is because in the analysis of spontaneous speech in Chapter 5 and the analysis of the read-aloud speech, the adjusted R^2s were much the same: 0.70 and 0.67, respectively (See 5.4).

6.4.2 Categorized pause[11]

The read-aloud speech of thirty-three Asian English learners was analyzed. The participants' first languages are Chinese, Tagalog, Korean Malay, Thai, and Japanese. Their average age was 23.0 years old with a SD of 3.6, and their average time studying English was 12.0 years with a SD of 4.1.

To operationalize learners' pause control, pauses are divided into three categories: sentential pauses, phrasal pauses, and within-phrase pauses. The number of pauses was counted in each category. The examples are presented in Table 6.7. Although Osada (2003) pointed out that the definition of a silent pause was a controversial issue, in this analysis, silent pauses are defined as silences beyond 100 ms, as examined in section 6.4.1. Multiple regression analysis was used to investigate the predictability of the evaluation score according to the speech characteristics: the criterion variable was the evaluation score and the predictor variables were the pruned syllables per second, the ratio of weak syllables to strong syllables, and the number of pauses in each category. A stepwise procedure was adopted, and the variables with a significance level of a partial correlation coefficient that was lower

[11] This section first appeared in 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.

than .05 were excluded.

The participants were divided into three groups according to their evaluation scores: high, mid, and low level. Each group consists of eleven participants. Figure 6.3 shows the differences in the average number of pauses in each pause category for each group. In sentential pauses, there was no difference among the three groups, but in phrasal pauses and within-phrase pauses, the learners at the higher level tended to make a smaller number of pauses than those at the lower level. This tendency was also found in the number of pruned syllables per second and the ratio of weak syllables to strong syllables, though in varying degrees: see 6.4.1. The correlation coefficients between these indices and the evaluation scores are shown in Table 6.8.

Table 6.7 Examples of pause categories

Category	Example
Sentential Pause	…when a traveler came along wrapped in a warm cloak. <P> They agreed that the…
	They agreed <P> that the one who succeeded in making…
Phrasal Pause	And at last <P> the North Wind gave up the attempt.
	…the North Wind was obliged to confess that <P> the Sun was the stronger of the two.
Within-phrase Pause	The <P> North Wind and the Sun were disputing…
	…the Sun shone out <P> warmly and immediately the traveler…

Note. <P> stands for silent pause.

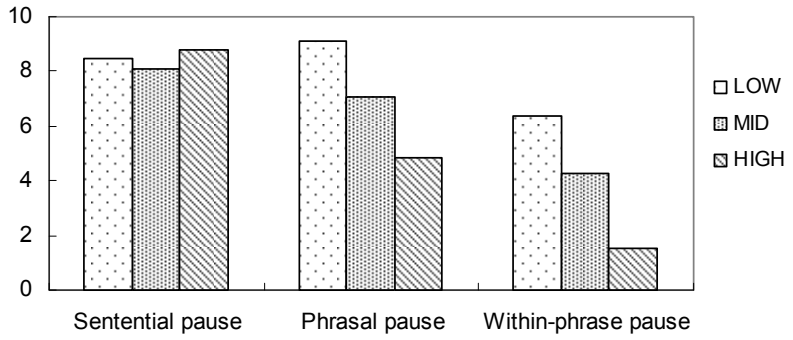

Figure 6.3 The Differences in Pause Control

Table 6.8 Correlation coefficients between scores and speech characteristics

	ES	PS	RWS	SP	PP	WP
ES	1	.83	-.60	-.15	-.64	-.53
PS		1	-.45	-.34	-.75	-.54
RWS			1	-.15	.44	.49
SP				1	-.32	.19
PP					1	.67
WP						1

Note.

ES: evaluation score

PS: pruned syllables per second

RWS: the ratio of weak syllables to strong ones

SP: sentential pauses

PP: phrasal pauses

WP: within-phrase pauses

As Table 6.8 indicates, the dominant predictor was pruned syllables per second, and the second leading predictor was the average number of phrasal pauses, but a negative high correlation was found between pruned syllables per second and the average number of phrasal pauses. There is no way of ascertaining the second and later predictors in this correlation

table. A negative substantial correlation was found between the average ratio of the weak syllables to the strong ones and the evaluation score. The range of the average ratio of the weak syllables to the strong ones is relatively narrow: between around .4 and 1.0 in the data (The average ratios of native English speakers are close to .5 or .4: Derwing, et al, 2004), and the lower scorers could not discriminate between the strong syllables and weak syllables in terms of the duration. That might be the reasons why we obtained the negative correlation here. In addition, the average ratio of the weak syllables to the strong ones obtained moderate positive correlations with the number of phrasal pauses and within-phrase pauses. The results indicate that pause making within sentences might relate to rhythmic control. In multiple regression analysis, two predictor variables were found to be statistically significant: pruned syllables per second and the ratio of weak syllables to strong syllables. The significance of the model was verified ($F_{(2, 32)} = 44.80$, $p < .01$). The model indicates that the evaluation scores can be predicted by these two variables to large extent (adjusted $R^2 = .73$). The equation below was obtained:

$$\hat{Y} = 1.35a_1 - 5.88a_2 - 1.29, \tag{3.7}$$

where a_1 is the pruned syllables per second, and a_2 is the ratio of weak syllables to strong syllables. This result means that among all the variables, these two indices independently predict the evaluation scores. The average number of phrasal pauses has a higher correlation with the evaluation score than the ratio of weak syllables to strong syllables, but the correlation of phrasal pauses with pruned syllables per second was so high that only a small portion of the score might be able to be predicted by the phrasal pauses alone.

This analysis revealed that pause control was one of the predictors of the evaluation scores given by human raters. However, pruned syllables per second were partially influenced by

the duration and number of pauses; therefore, the index of pause control could not be a statistically significant predictor.

6.4.3 Vowel discrimination[12]

In the analyses so far, the focus of the examination of learners' speech characteristics was on prosodic features, although the evaluation items of the read-aloud speech included speech characteristics other than prosodic features such as loudness, pitch, and the quality of segmental sounds. The acquisition of segmental sounds is a controversial issue in the field of second language acquisition. Many studies have investigated the relationship between foreign accentedness and a learners' background: the arrival time to the country of the target language, the first language, and the motivation to study the target language (e.g. Flege, 1987; Piske, Mackay, & Flege, 2001). Other speech features of learners, such as pitch range and intonation, are considered to be the index of the attainment of the target language. These features can also be evaluated in read-aloud speech. Although the learners' prosodic features are dominant predictors of the evaluation of read-aloud speech, as shown in the analyses so far, it is natural to consider features other than prosodic ones. This section explores the relationship between vowel discrimination rate and evaluation scores, focusing on the difference in vowel systems between the target language and the first language (Japanese).

Speech spoken by thirty-eight Japanese learners of English was analyzed. Their average age was 20.4 years old with a SD of 2.1, and their average time spent studying English was 9.3 years with a SD of 2.8. They were divided into three groups: high, mid, and low levels, according to their evaluation score. Data from five participants were randomly selected from each group and acoustically measured. This decision was made because only five

[12] This section was first appeared as Kitagawa, A., Kondo, Y., & Nakano, M. (2007). Does vowel quality matter? *Proceedings of 12th Conference of Pan-Pacific Association of Applied Linguistics*, 224-227.

Japanese learners belonged to the high level.

Taking account of the differences and the similarities between the Japanese and English vowel systems, three pairs of vowels were chosen: /ɪ/ and /i/, /u/ and /ʊ/, and /æ/ and /ʌ/. The Japanese have five vowels in quality, and each vowel can be both long and short. On the other hand, English has a much larger vowel system with about eleven or twelve vowels. Assuming the interlanguage transfer of vowels, the Japanese tend to produce two or more distinct English vowels with one Japanese vowel. For this reason, the three targeted pairs of English vowels are considered to be less distinguishable for Japanese English learners.

Acoustic measurements were performed with the acoustic analysis software, Praat. First, segmentation was provided to each speech. Then, the F1 and the F2 of the target vowel were measured at the point that was considered to be under the least influence from the sounds adjacent to it. This point was visually defined by hand with the help of formant tracks from F1 to F5. The words including the target vowel were as follows (The number in the brackets indicates how many times the word was spoken): agreed [1], succeed [1], and immediately [1] for /i/; wind [4], which [1], and considered [1] for /ɪ/; disputing [1], blew [2], and two [1] for /u/; should [1], could [1], and took [1] for /ʊ/; traveler [4], wrapped [1], and last [1] for /æ/; and sun [3], one [1], other [1], and up [1] for /ʌ/. If a word was repeated, the F1 and F2 values were averaged; as a result, the data of each target vowel was obtained from three or four different words respectively. Whether or not each speaker differentiated each vowel from another was examined as a physical reference of the achievement in vowel quality. A statistical test, a discriminant analysis, was conducted to investigate the speakers' achievements in classifying these six vowels.

The F1 and F2 values are presented in Figures 6.4, 6.5, and 6.6. F1 values are plotted on the y-axis, and F2 values are plotted on the x-axis. Examples were selected from each level. As the level dips from the high to low, the target pairs of vowels become indistinguishable

based on their F1 and F2 values. For example, /i/ (♦), /ʊ/ (▲), /æ/ (*), and /ʌ/ (●) are distinguishable in the production by a high-levels examinee (Figure 6.4), but in the production of a low-level examinee, all vowels are indistinguishable by F1 and F2.

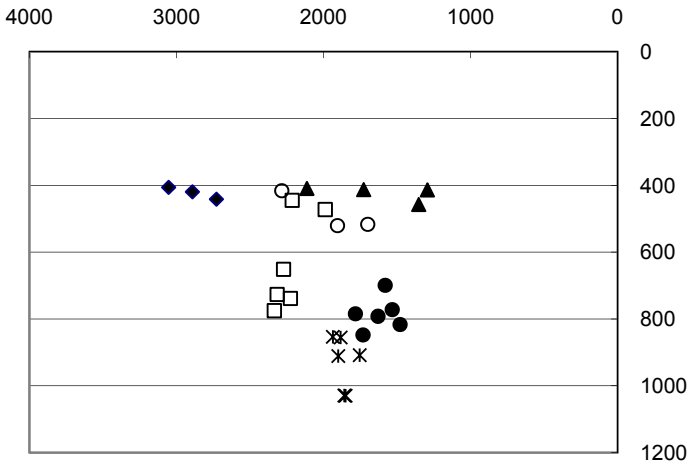

Figure 6.4 F1 and F2 Values of a High-level Examinee

Note. ♦, □, ▲, ○, *, and ● indicates /i/, /ɪ/, /ʊ/, /u/, /æ/, and /ʌ/ respectively.

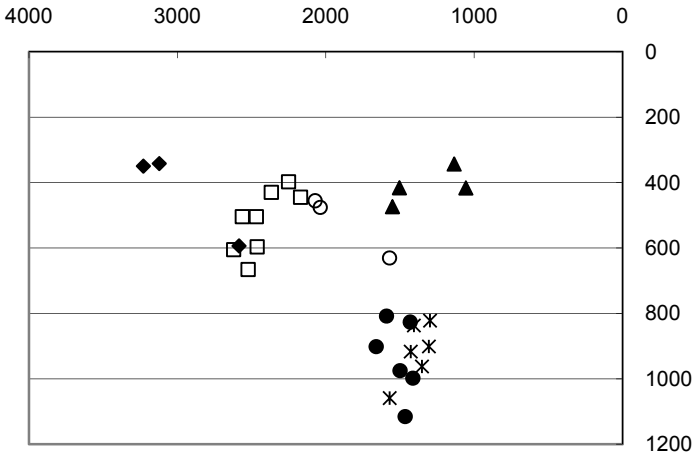

Figure 6.5 F1 and F2 Values of a Mid-level Examinee

Note. ♦, □, ▲, ○, *, and ● indicates /i/, /ɪ/, /ʊ/, /u/, /æ/, and /ʌ/ respectively.

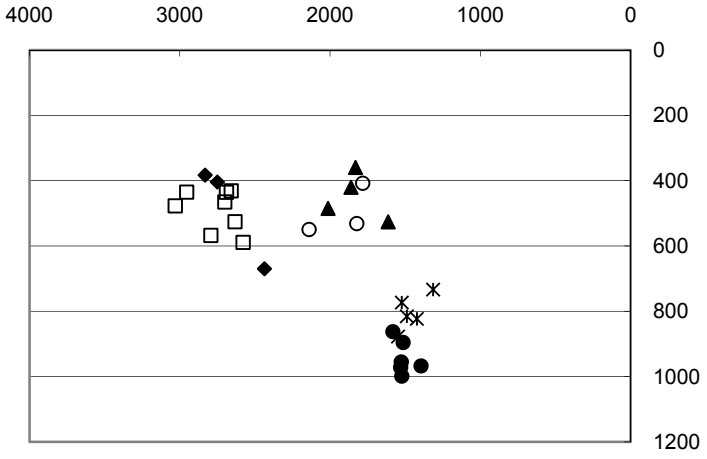

Figure 6.6 F1 and F2 Values of a Low-level Examinee

Note. ♦, □, ▲, ○, *, and ● indicates /i/, /ɪ/, /ʊ/, /u/, /æ/, and /ʌ/ respectively.

Tables 6.9, 6.10, and 6.11 show the results of the linear discriminant analysis with cross validation between every possible pair for each group according to overall pronunciation proficiency. The correct classification rates between the target pairs (i.e., /i/ and /ɪ/, /ʊ/ and /u/, and /æ/ and /ʌ/, whose counterparts in Japanese are /i/, /u/, and /a/, respectively) were fairly low for all the groups; however, the speakers in the high-level group tended to succeed in classifying /i/ and /ɪ/ (86.7%).

Table 6.9 The ratio of correct classifications (high-level)

	/ɪ/	/u/	/ʊ/	/æ/	/ʌ/
/i/	86.7	90.0	90.0	100	100
/ɪ/	-	70.0	80.0	100	100
/u/	-	-	63.3	100	100
/ʊ/	-	-	-	96.7	100
/æ/	-	-	-	-	62.9

Table 6.10 The ratio of correct classifications (mid-level)

	/ɪ/	/u/	/ʊ/	/æ/	/ʌ/
/i/	51.4	80.0	83.6	100	100
/ɪ/	-	75.9	75.9	96.4	97.1
/u/	-	-	63.3	96.6	94.3
/ʊ/	-	-	-	96.6	91.4
/æ/	-	-	-	-	53.7

Table 6.11 The ratio of correct classifications (low-level)

	/ɪ/	/u/	/ʊ/	/æ/	/ʌ/
/i/	43.3	90.0	76.7	100	97.1
/ɪ/	-	90.0	83.3	100	100
/u/	-	-	56.7	93.1	88.6
/ʊ/	-	-	-	93.1	91.4
/æ/	-	-	-	-	55.9

Although these results indicate the Japanese learners' vowel pronunciation features, they cannot be the factor that distinguishes the learners' read-aloud speech levels. Although the examinees' discrimination rate of /i/ and /ɪ/ at the high-levels is very high, especially when compared with the examinees in the other two levels, the discrimination rates are not ideally suited for the other vowel pairs. Ideally, the data would have shown that the high-level examinees obtained higher discrimination rates, the mid-level obtained moderate rates, and the low-level obtained the lowest rates.

6.4.4 Vowel reduction[13]

The speech of thirty-eight Japanese English learners was analyzed in terms of vowel reduction. They were divided into three groups, high, mid, and low levels, according to their evaluation score. Data from five participants were randomly selected from each group and acoustically measured. This decision was made because only five Japanese learners belonged to the high level. This data is the same data that was used in section 6.4.3.

English reduced vowels in unstressed syllables belong to one of the three vowels: /ɪ/, /ʊ/, and /ə/. According to Roach (2000), these unstressed vowels, or weak syllables, are likely to be shorter in duration and have lower intensity and different qualities than stressed (strong) syllables. On the contrary, in the Japanese prosodic system, a change of F0 is required in order to realize the accent and the long-short contrast in sound length to achieve the mora duration. Furthermore, variations in intensity and vowel quality are not necessary in Japanese phonology. In consideration of these differences between the English and Japanese phonological systems, it is hypothesized that these features of reduced English vowels can be adopted as the indices that categorize a learner's proficiency when reading aloud.

The data were grouped into three levels, high, mid, and low, and five data sets were randomly selected from each group and were acoustically measured; each group was comprised of two males and three females.

This analysis only focused on /ə/ as the target vowel, and this vowel was analyzed with the acoustic analysis software, Praat. The test-tokens were /ə/ in the words "attempt," "around," "agreed," "along," "considered," "confess," and "obliged." The underlined vowels are supposed to be produced as /ə/. First, each participant's speech was segmented. Then, the target features, F0, duration, intensity, and F1 and F2 were measured. With regard to the

[13] This section first appeared in Kitagawa, A., & Kondo, Y. (2008). Reduction of vowels by Japanese learners of English. *Proceedings of 13th Conference of Pan-Pacific Association of Applied Linguistics*, 227-230.

first three properties, the values of stressed vowels within the same word (the bolded vowels) were also analyzed in order to calculate their relative values, and the ratios of unstressed vowels to stressed vowels were obtained. The F1 and the F2 of the target vowels were measured at the point that was considered to be under the least influence from adjacent sounds. This point was visually defined by hand with the help of formant tracks from F1 to F5. Additionally, these values were normalized in order to compare the data across the speakers based on Guion's method (Guion, 2003). In Guion's normalization, first, one speaker's F3 value for /æ/ is taken as a norm, because F3 is commonly recognized as a reflection of vocal tract length. Second, this norm F3 value is divided by the mean F3 values for /æ/ produced by each speaker, and the factor for each speaker is calculated. Finally, the F1 and F2 values are multiplied by this respective factor. In this study for instance, speaker A's average F3 value for /æ/, 2696.42, was taken as a norm. Given speaker B's average F3 value for /æ/, 2275.35, the factor for speaker B was 1.185 (2696.42 divided by 2275.35). Then the normalized F1 and F2 values of speaker B were obtained by multiplying each with 1.185. The average F3 values for /æ/ and the factors of all speakers are shown in Table 6.12, 6.13, and 6.14. These formant values were transformed to mel scale in order to examine the perceptual vowel quality. Mel scale is a perceptual scale of sound pitch. The difference in mel scale indicates the difference of which a human being senses sound pitch. This scale is defined as:

$$\mathrm{Mel(f)}2595\log_{10}(1+\frac{\mathrm{f}}{700}) \tag{6.1}$$

where f is hertz (Young, Evermann, Gales, Hain, Kershaw, Liu, Moore, Odell, Ollason, Povey, Valtchev, and Woodland., 2006).

Table 6.12 The average F3 values for /æ/ and the factors of speakers at high-level

	Speaker A	Speaker B	Speaker C	Speaker D	Speaker E
Ave. F3 value	2696.42	2275.35	3035.17	2166.65	2084.64
Factor	-	1.185	0.888	1.244	1.293

Table 6.13 The average F3 values for /æ/ and the factors of speakers at mid-level

	Speaker F	Speaker G	Speaker H	Speaker I	Speaker J
Ave. F3 value	2323.63	2148.52	2739.49	3127.39	2339.38
Factor	1.160	1.255	0.984	0.862	1.152

Table 6.14 The average F3 values for /æ/ and the factors of speakers at low-level

	Speaker K	Speaker L	Speaker M	Speaker N	Speaker O
Ave. F3 value	2616.11	2430.06	2741.08	2668.66	2226.15
Factor	1.030	1.109	0.983	1.101	1.211

In Figures 6.7, 6.8, and 6.9, F1 and F2 values are plotted depending on the orthographic spelling. As far as the visual observation goes, it can be noted that unstressed vowels were more centralized in their quality for the high-level and mid-level groups than the low-level group. Reduced vowels produced by the low-level group were more separated in the vowel space according to the spelling, divided into "a" space and "o" space. The fact that the mid-level group probably performed better than the high-level group may propose that the accuracy of vowel quality reduction does not matter after a certain degree of accuracy is satisfied.

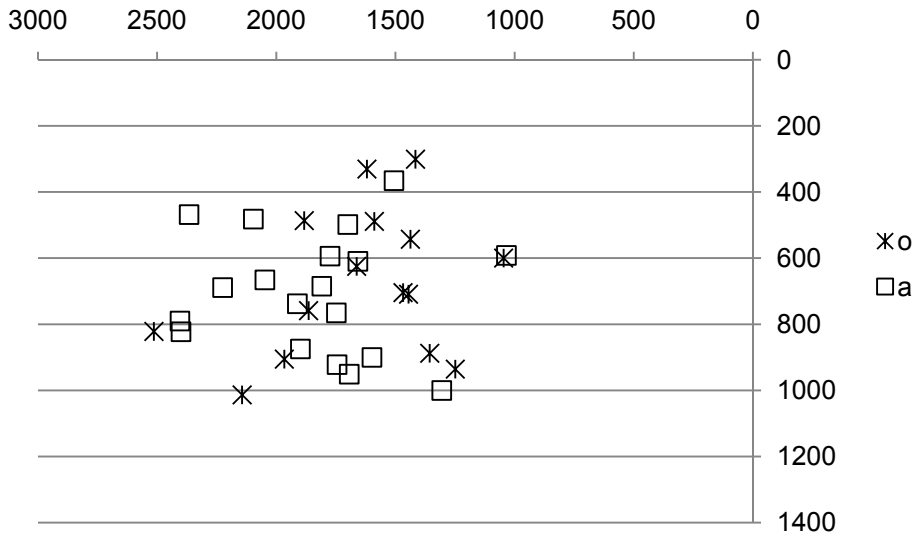

Figure 6.7 F1 and F2 of the Reduced Vowels at High-level

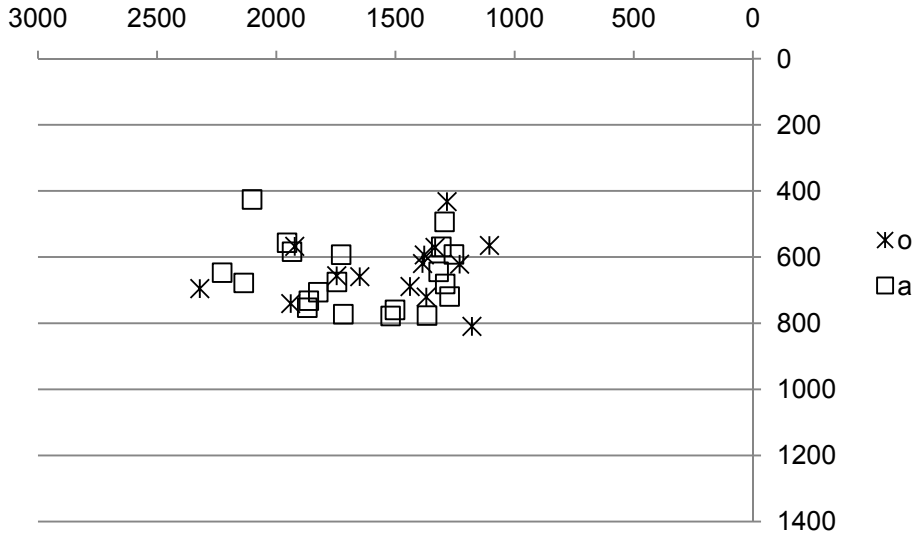

Figure 6.8 F1 and F2 of the Reduced Vowels at Mid-level

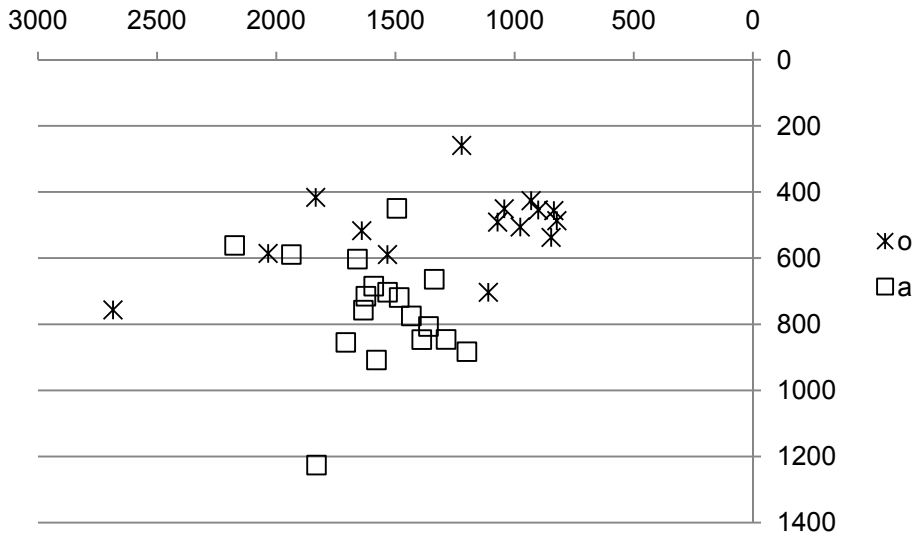

Figure 6.9 F1 and F2 of the Reduced Vowels at low-level

Table 6.15 Mean and standard deviation of intensity

	High	Mid	Low
Mean	-2.02	-2.09	-1.51
S.D.	3.99	3.41	3.57

Table 6.16 Mean and standard deviation of F0

	High	Mid	Low
Mean	.92	.91	.92
S.D.	.07	.08	.10

Table 6.17 Mean and standard deviation of duration

	High	Mid	Low
Mean	.47	.52	.57
S.D.	.29	.25	.27

The means and standard deviations of intensity, F0, and duration are shown in Tables 6.15, 6.16, and 6.17. For intensity, the log ratios calculated by subtracting the intensity of stressed vowels from that of unstressed vowels are shown in Table 6.15. If the value was less than 0, the intensity of the unstressed vowels was successfully lower. For F0 and duration, the ratio of unstressed syllables to stressed syllables is shown in Table 6.16. If the value is less than 1, the F0 of the unstressed vowel is successfully lower and the duration is successfully shorter.

As shown in Table 6.15, the intensity of the unstressed vowels produced by the learners in all groups was weaker than stressed vowels, but there was no significant difference between groups. This was also the case with the analysis of F0, where little difference was found among the groups. The results indicate that these features cannot be used as a predictor of the evaluation score when compared with the duration shown in Table 6.17.

However, as Table 6.17 shows, the analysis of the duration reveals a remarkable difference among the groups. The results of the analysis shown here represent the ratio of unstressed vowels to stressed vowels. This analysis is considered to be a detailed analysis of the index of rhythm that was adopted in the analysis of prosodic features reported in 6.4.1. In the analysis of prosodic features, as with the index of rhythm, the ratio of unstressed syllables to stressed syllables in terms of the time duration required to complete the passage was adopted, and it was verified as one of the statistically significant predictors of the evaluation score. The results of this detailed analysis of unstressed vowels also confirm that time duration of stressed syllables and unstressed syllables is one of the important factors that predicts the evaluation score for speech that is read aloud.

6.4.5 Loudness, pitch, and pronunciation errors

The read-aloud speech of eight Japanese, eight Korean, seven Chinese, three Thais, three Filipinos, and two Malaysians was analyzed. There were thirty-one participants, their

average age was 23.19 years old with a SD of 3.63, and their average time spent studying English was 13.48 years with a SD of 4.84. The evaluation scores were used as the criterion variable, and the speech characteristics discussed below were used in multiple regression analysis. The data were the same as the data used in 6.4.1

The index of loudness of voice was operationalized as the largest value of loudness (dB) while reading aloud. The index of sound pitch was operationalized as the range of pitch (Hz) while reading aloud. As for the indices of elision and epenthesis, through the observation of the data, few elisions and epentheses were found. Thus, as the index of pronunciation error, the sum of the number of elisions, epentheses, and replacements were examined. An example of an elision found in this analysis is the deletion of consonants (e.g., in the word "succeeded" /səksidɪd/ → /səsidɪd/). There was no deletion of vowels found in the present data. Examples of epenthesis were observed at the end of words that end with plosive sounds such as "first," "wind," and "fold." Replacement was defined as the replacement of vowels and consonants. Examples of the pronunciation errors found in this analysis were the replacement of words (e.g., cloak → coat), the replacement of vowels (e.g., in the word "obliged" /aɪ/ → /ɪ/), and the replacement of consonants (e.g., in the word "first" /f/ → /p/). Through the observation of the data, there was no addition of consonants. The measurements of these features were done with the acoustic analysis software, Wavesurfer (Sjölander and Beskow, 2000). Regarding the measurement of pronunciation errors, each point was visually defined by hand with the help of formant tracks from F1 to F5.

Table 6.18 shows the mean and the standard deviation of the indices of loudness, pitch, and pronunciation error. Table 6.19 shows the correlation coefficients between the evaluation score, loudness, pitch, and pronunciation error.

Table 6.18 The mean and standard deviation of loudness, pitch, and pronunciation error

Feature	*M*	*S.D.*
Loudness	65.72	5.13
Pitch	238.91	47.67
Pronunciation error	2.60	3.00

Note. n = 33.

According to the correlation table below, the correlation coefficients between the evaluation score and the pronunciation error index is fairly high, but that index substantially correlates with the indices of speech rate and rhythm (–.45 and .62, respectively), which were demonstrated to be statistically significant predictors of the evaluation score in the analysis of speech timing control characteristics.

Table 6.19 The correlation coefficients between evaluation score, loudness, pitch, and pronunciation error

	ES	PS	RWS	PT	PE	LN
ES	1	.75	-.46	.37	-.68	.33
PS		1	-.05	.21	-.45	.06
RWS			1	-.28	.62	-.38
PT				1	-.29	.24
PE					1	-.04
LN						1

Note. n = 33

ES: evaluation score

PS: pruned syllable per second

PT: pitch

PE: pronunciation error

RWS: the ratio of weak syllable to the strong one LN: loudness

In multiple regression analysis (stepwise method) where the criterion variable was the evaluation score and the predictor variables were pruned syllable per second, the ratio of the weak syllables to the strong ones, pitch, the number of pronunciation error, and loudness, the latter four variables were not found to be statistically significant predictors of the evaluation score.

6.5 Summary and discussion

In this chapter, the evaluation of the read-aloud speech of L2 learners was analyzed based on MFRA. Neither raters nor evaluation items exceeded the standards for consistency. Furthermore, though a little difference was found in the severity of the raters and the difficulty of the items, the ranges of infit statistics for both raters and items was found to be narrower than those observed in the evaluation of spontaneous speech. These results are attributed to the rater training, the fact that the raters became familiar with the evaluation of recorded speech on the website, and the fact that the variety of the speech was limited because it was read from a text. From this point of view, we can suggest that the read-aloud speech was more adapted for the performance assessment than the spontaneous speech in terms of the evaluation of prosodic features.

The aim of the correlation studies in this chapter was to examine the predictability of the speech evaluation scores by using the learners' speech characteristics that could be objectively measured in the read-aloud speech. In the spontaneous speech analysis, the two speech characteristics of speech rate and number of filled pauses were verified as statistically significant predictors of evaluation scores. This indicates that the dominant factors in L2 speech evaluation are the speech timing control characteristics. The adjusted R^2s obtained in

the analysis of the speech-timing control characteristics and the categorized pauses were .67 (see 6.4.1) and .73 (see 6.4.2), respectively. These results led us to conclude that the human rating in the evaluation of read-aloud speech can be predicted to large extent by the two previously mentioned speech characteristics. The adjusted R^2 obtained in the analysis of the spontaneous speech was .70 (see 5.4). Much the same R^2s were obtained in these studies. As the results of Cucchiarini, Strik, and Boves (2002) indicate, in L2 speech performance assessments, the speech rate index is a dominant predictor for human rating, but the correlation between speech rate and rating decrease if speech includes syntactic and lexical variety.

Five pilot studies were conducted: analyses of speech timing control characteristics, pause control, vowel discrimination, vowel reduction, and loudness, pitch, and pronunciation error. The analysis of speech timing control characteristics revealed that two characteristics—pruned syllables per second and the average ratio of weak syllables to the strong syllables—were significant predictors of the evaluation scores. In the other four analyses, learners' pause control, vowel discrimination, reduced vowels, loudness, pitch, and pronunciation errors were examined. Although several indices were found to substantially correlate with the evaluation scores, the indices that were examined in the analyses were not statistically significant predictors of the evaluation scores. In the analysis of vowel discrimination, all three groups (high, mid, and low levels) showed a similar tendency to unsuccessfully separate the six distinct vowels. Accordingly, what can be inferred from this analysis is that overall pronunciation proficiency is less related to vowel quality. This result conflicts with that of Cucchiarini, Strik, and Boves (2000b). In their study, although pronunciation was not a good predictor of human rating, it moderately correlated with the rating. This is ascribed to the differences in the target language between Cucchiarini, et al. (2000b) and the present study, namely Dutch and English. The raters of English should have

a wider acceptable range of pronunciation than Dutch raters. English is currently used, learned, and taught throughout Asia. Our raters who are L2 users of English are able to understand the context of learning English in Asia. Their learners do not need to acquire native-like pronunciation: see 2.3.

The analysis of pause control revealed that pause insertion is one of the clues that raters use to evaluate learners' speech. However, although the speech rate index used in this analysis was partially influenced by the number and duration of pauses, the pause control index could not be a statistically significant predictor.

In the analysis of reduced vowels, the fundamental frequency, intensity, and duration of reduced vowels were examined and compared with similar factors in stressed vowels. Although some limitations need to be taken into account when considering the results because the number of the informants were fairly small in this analysis, the only feature that differentiated the groups of learners (high, mid, and low) was the time duration of reduced vowels, which confirmed the ability of the ratio of weak syllables to strong syllables to predict the evaluation scores.

In the analysis of loudness, pitch, and pronunciation errors, the number of the pronunciation error defined substantially correlates with the evaluation scores, but this index was substantially correlated with the index of rhythm, the ratio of weak syllables to strong syllables. In multiple regression analysis, the number of the pronunciation error was not identified as a significant predictor of the evaluation score.

Through the five pilot studies on the relationship between the evaluation scores and L2 learners' speech characteristics, the two statistically significant predictors of the evaluation score were the indices of speech rate and stress timing. However, the results of the studies do not simply imply the significance of these two speech characteristics. Other characteristics of read-aloud speech were also found to be predictors of the evaluation scores.

In the analysis of speech timing control characteristics, pruned syllables per second obtained the highest correlation coefficient with evaluation scores; however, substantial correlations were found among the other indices: the number of silent pauses, duration of silent pauses, mean length of run, and the number of unnecessary syllables (their correlation coefficients were –.62, –.56, .67, and –.51, respectively: See Table 6.5). Furthermore, in the analysis of categorized pauses, the average number of phrasal pauses and within-phrase pauses was substantially correlated with evaluation scores (–.64 and –.53, respectively: See Table 6.8). Considering the definitions of these indices, correlations can be found among them. Using the number of silent pauses as an example, an increase in the number of pauses makes pruned syllables per second drop. A similar sort of relationship is found among the other characteristics. Taking the results of the studies into account, the index of speech rate adopted in the present studies is considered to be one of the representative indices of prosodic features that can predict the evaluation of L2 learners' read-aloud speech. Because these characteristics were found to correlate with the pruned syllables per second, they were unqualified to be good predictors of the evaluation scores in the multiple regression analysis that included pruned syllables per second as a predictor variable.

In the analyses of vowel discrimination and reduced vowels, though some tendency was found in vowel discrimination, no good predictors of evaluation scores were discovered except the time duration of reduced vowels. These results indicate that the vowel quality is not a key characteristic in the overall evaluation of L2 speech. This confirms that stress timing control is a good predictor of the evaluation of prosody and pronunciation in read-aloud speech. The index used in the analysis of speech timing control characteristics—the ratio of weak syllables to strong syllables—was found to be a statistically significant predictor of evaluation scores. These results imply the prominence of suprasegmental features, rather than segmental features. However, in the results of the

analysis of loudness, pitch, and pronunciation errors, the number of pronunciation errors was found to be substantially correlated with evaluation scores. This seems to contradict the results of the analysis of vowel discrimination and reduced vowels. However, these two results are different with regard to level of analysis. In the analysis of vowel discrimination and reduced vowels, the target features were continuous variables. For example, the vowel discrimination rates were calculated based on the F1 and the F2 of target vowels. In the analysis of pronunciation errors, on the other hand, the target features are dichotomous variables. The number of pronunciation errors was the sum of the replacement, deletion, and epenthesis that individual learners performed as they read aloud. The difference of the target variables was considered to cause a contradiction in the interpretation of the results of these two studies. This can be interpreted as an inability of the raters to pay attention to vowel quality during the evaluation of pronunciation in read-aloud speech; however, they were able to use pronunciation errors as clues that suggest learners' proficiency levels. Furthermore, as indicated by the analysis of loudness, pitch, and pronunciation errors, the number of pronunciation errors is substantially correlated with the indices of speech rate and rhythm (–.45 and .62, respectively: See Table 6.19). Furthermore, in multiple regression analysis, the number of pronunciation errors was not found to be a statistically significant predictors of the evaluation scores. The number of pronunciation errors is conceptually different from the indices of speech rate and rhythm, but there may be a general tendency for L2 learners who make few pronunciation errors while reading aloud to speak faster and have better stress timing control.

7 Asian English speech database

7.1 Introduction

To fully understand the speech characteristics of second language (L2) English learners in Asia, it is important to create an L2 speech database from the view point of Asian English language education. Furthermore, in the analysis of L2 speech by a speech recognizer, it is essential to collect a great deal of L2 speech data, because current speech technology is adopting a statistic approach to speech recognition (e.g. Hidden Markov Model). This chapter describes the design of an Asian English speech database, which is a collection of the read-aloud speech of L2 English learners, and the speeches are evaluated by human raters. The purposes for constructing this database are to shed light on the rating behaviors of L2 raters and to provide data for creating an automatic L2 speech evaluation system. Now this project broadens its candidates to include repetition, discourse completion task, quasi-spontaneous speech, and spontaneous speech. This chapter is organized in the following manner. First, works related to such a database are reviewed. Second, the procedure for speech recognition that is adopted in this study is introduced. Third, the design of the database is illustrated. Lastly, several issues in non-native speech databases are discussed.

7.2 Related work to the study

Raab, Gruhn, and Noeth (2008) reviewed non-native speech databases. In their article, thirty-three non-native speech databases were described, and information on these databases is updated on a companion website (Wikipedia). The databases listed in Raab et al. (2008) include several non-native databases that were created for different purposes. Almost all of the databases are for the training of speech recognizers. Among them, however, there are

only three databases with a proficiency rating; two of them are databases of English learners, and the other is a database of Japanese learners. Although this article does not contain all existing databases, as the authors mention, a small number of the speech databases of English learners are available. Moreover, in the speech databases of English learners with proficiency ratings that are listed in Raab et al. (2008), there is little attention paid to the rating procedure.

In Minematsu, Tomiyama, Yoshimoto, Shimizu, Nakagawa, Dantsuji, and Makino (2003), a non-native speech database listed in Raab et al. (2008), they mention that their raters did not discuss their rating procedures, evaluation items, or their criteria before their evaluation. However, it is unlikely that raters who are not in agreement about the evaluation judge learners' performance based on their English language education where English is learned as a second or foreign language. Furthermore, reliable evaluation scores are one of the essential requirements for the construction of an automatic speech evaluation system because recent studies on automatic scoring use an approach that predicts evaluation scores of human raters using the speech characteristics of learners' speech. The examination of rating procedures and the reliability of evaluation scores is a crucial step in the construction of a non-native speech database that makes use of a proficiency rating.

7.3 Speech recognizer[14]

To measure learners' speech characteristics, the system adopted the Hidden Markov Model Toolkit (HTK). HTK is a tool for Hidden Markov Model (HMM) that has been optimized for speech recognition (see Young, et al, 2006:2-13, for the details of the use of HMM in speech recognition). The procedure of model training in HTK is depicted in Figure 7.1.

[14] This section is based on Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *HTK book*. Cambridge University Engineering Department.

Firstly, HTK codes the raw speech waveforms into sequences of feature. In this study, Mel Frequency Cepstral Coefficients were used. In the model training, because HTK requires prototype HMM, text labels for the speech data, and a pronunciation dictionary (See Appendix E) were created. The phonetic descriptions were completed, based on Jones (2003). The phonetic symbol table for the pronunciation dictionary is shown in Appendix F. For the initial training, the speech data from TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo, Lamel, Fisher, Fiscus, Pallett, Dahlgren, and Zue: 1993) were used to train the HMM. To adapt the model to English spoken by Asian learners, the speech data of the read-aloud speech of 101 Asian English learners were used. In the process of adopting and training the model, HTK phone-aligned the target speech data based on the order of occurrence of phones by referring to the text labels and the pronunciation dictionary. HTK must run through model training several times to create robust HMMs. A gender-independent HMM recognizer was bootstrapped to native speech data and was trained by using non-native speech data. The initial speech data were .wav format, 24 kHz sampling rate, in stereo, and were then transformed into 16 kHz monaural files.

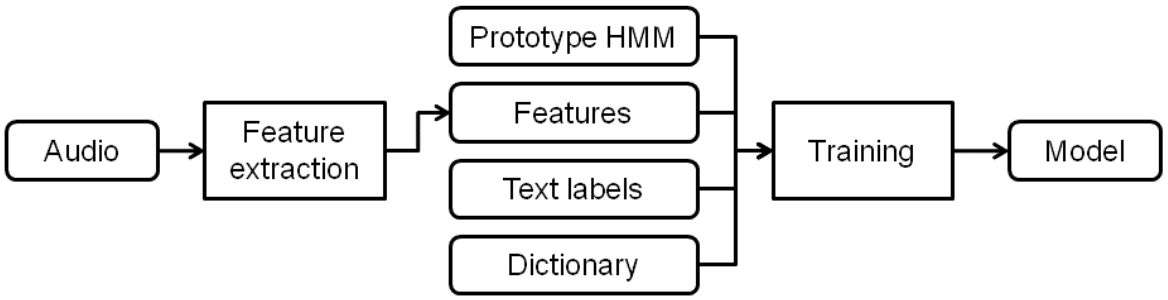

Figure 7.1 The Model Training Procedure in HTK

Figure 7.2 shows the procedure of the forced alignment. In this procedure, HTK phone-aligns the speech data, based on the corresponding text labels. Phonetic time

alignments were generated for the speech data using the Viterbi algorithm with the native English model (TIMIT Acoustic-Phonetic Continuous Speech Corpus) trained with the speech data of Asian learners of English. Through this process, phone-aligned speech data are obtained. Figure 7.3 shows an example of phone-aligned speech data (The phone-aligned speech data of "The North Wind and the Sun").

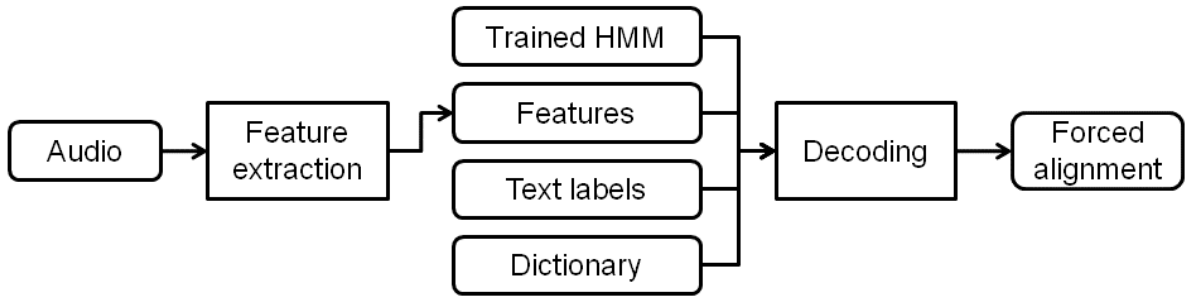

Figure 7.2 The Forced Alignment Procedure in HTK

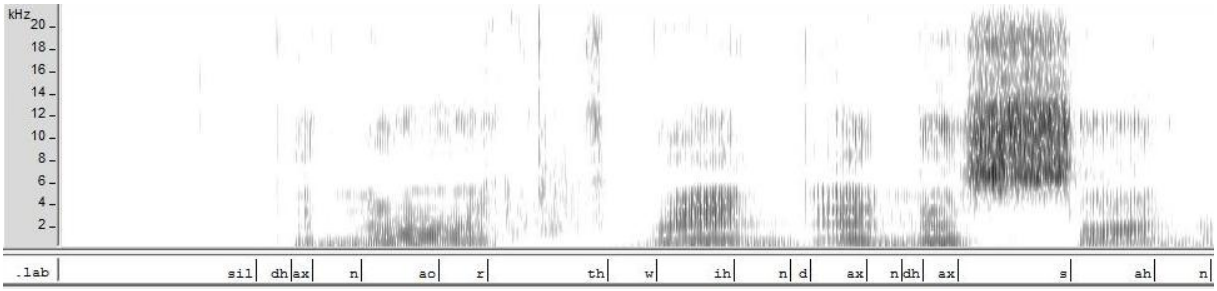

Figure 7.3 An Example of Phone-aligned Speech with its Spectrogram

7.4 Database design

7.4.1 Material

The reading text that the speakers read aloud was a fable from Aesop, "The North Wind and the Sun," which is so famous that students at the university level should know it. This passage was also used in the National Institute of Education Singapore (NIE) corpus (Deterding and Ling, 2005), and is used in the phonetic description of the International

Phonetic Association. The passage below is used in this database.

> 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.

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 Jones, 2003).

7.4.2 Speakers

One hundred and one Asian English learners participated in the recording of the passage. There were forty Japanese participants, seventeen Chinese, nineteen Korean, six Filipino, ten Thai, four Vietnamese, four Cambodians, and one Indonesian. They were all either undergraduate or graduate students. For all speakers, the followings are provided in the database:

- Speaker ID
- Age
- Sex
- Native language
- The period for which a speaker studies English
- The hours when a speaker uses English in a week
- The period that a speaker studied in the English spoken countries
- Evaluation score
- Transcription
- Read-aloud speech

7.4.3 Recording procedure

All the recordings were made in soundproof rooms located in the universities that the speakers belonged to. Informed consent was obtained from all speakers in this database to use their speech for only academic purposes. The sheet for the informed consent can be found in Appendix G. The speakers were individually called into the room and given recording instructions. Their speech was digitally recorded by using a Roland R-09 and a condenser microphone, SONY ECM-MS957. In the process of recording, first, the participants introduced themselves to an interviewer in order to check the volume of the microphone, and second, the speakers read the passage silently to understand its contents. Then they read the passage out loud, and the read-aloud speech was digitally recorded. 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.

7.4.4 Orthogonal transcription

All speech material was checked and orthographically transcribed. Repetitions, insertions of words that did not appear in the text, and self-corrections were transcribed exactly as pronounced. Any pronunciation errors were also transcribed as pronounced. Examples of the pronunciation errors found in this database are deletion of consonants (e.g., in the word "succeeded" /səksidɪd/ → /səsidɪd/), epenthesis (adding vowels to the end of words that end with plosive sounds such as "first," "wind," and "fold"), and the replacement of vowels or consonants (e.g. /aɪ/ → /ɪ/ in the word "obliged," and /f/ → /p/ in the word "first").

7.4.5 Raters and rating procedure

Five Japanese language teachers evaluated the speeches using fourteen evaluation items on a 6-point Likert scale. The raters received rater training where they discussed the characteristics of the learners' speech through watching a video provided by Council of Europe (North and Hughes, 2003): See Chapter 4. This video depicted learners that were divided into six levels according to the Common European Framework of References (CEFR). Unreliable raters were excluded based on Multifaceted Rasch Analysis (MFRA). The raters evaluated the recorded spontaneous speech of seventy three Asian English learners. The evaluation items were selected from Yashiro, Araki, Higuchi, Yamamoto, and Komissarov (2001), and each item was thoroughly reviewed in order to make sure that the item was suitable for the evaluation of unprepared L2 speech: see 5.3. The items were also scrutinized based on MFRA. Then, evaluation items suitable for the evaluation of read-aloud speech were selected. Using these items, the raters evaluated the speeches individually using the evaluation website. The scores credited to the read-aloud speech represented a learner's ability as estimated by MFRA: see 6.3. The estimated ability resulting from MFRA is a variable calculated, based on raw scores, and so the correlation of

the estimated ability using raw scores is fairly high, but it was calculated considering the severity of the raters and the difficulty of evaluation items.

7.4.6 Example

In addition to the information on a speaker's background and evaluation score, the sound files of the speech, and the transcription, the information from the phone-alignment is provided in this database. Below is an example of a phone-aligned utterance created by a speaker saying "The North Wind and the Sun."

0	500000	sil
500000	1500000	dh
1500000	3000000	ih
3000000	12500000	sil
12500000	13800000	n
13800000	14500000	ao
14500000	15500000	r
15500000	16400000	th
16400000	17700000	w
17700000	18600000	ih
18600000	19200000	n
19200000	20000000	d
20000000	20600000	ax
20600000	21100000	n
21100000	21400000	dh
21400000	21900000	ax

21900000	23500000	s
23500000	24900000	ah
24900000	25700000	n

The first and the second columns indicate the beginning and ending times when the phones in the third columns were uttered. The figures in the first and second columns were expressed in ten millionths of a second. The results of the phone alignment by the speech recognizer were visually checked using an acoustic analysis software, Wavesurfer (Sjölander, K. and Beskow, J., 2006) and were manually modified.

7.5 Issues in a non-native speech database

Constructing a non-native speech database involves additional difficulties that are not found in the native speech corpus. Schaden and Jekosch (2006) mentioned the issues that arise from elicitation methods. Speakers feel some anxiety when they are asked to read a passage in a foreign language, and moreover, they feel embarrassed to some extent when their speech is recorded, even in cases where it is not related to any sort of L2 proficiency test. In the case of the present database, because the author and his co-researchers who are L2 users, visited universities that the speakers belonged to in order to obtain the recording, it was apparent to the speakers that the recording was not involved in any sort of L2 proficiency test. Furthermore, the speakers might have been more relaxed when recording with an L2 user, rather than recording with a native English speaker. However, this problem is fairly difficult to solve in controlled recording settings. The second issue in non-native speech databases that Schaden and Jekosch (ibid) mentioned is the difference between “laboratory” and “natural” settings. Although this is also a problem in the case of collecting native speakers’

speech, L2 speakers tend to pay much more attention to the language that they use than native speakers do. This self-monitoring hinders the collection of "natural" non-native speech.

However, these two problems are not a problem in the present study. The purpose of constructing the present database is to build an automatic speech evaluation system. Speech data required in the database is the speech in a test situation.

7.6 Final remarks

The Asian English speech database constructed in this study exhibits three distinguishing features. Firstly, the speakers' language backgrounds are limited to Asian languages. English is now taught, learnt, and widely used throughout Asia. This database is meant to be of assistance in studying the speech characteristics of Asian learners of English. Secondly, the passage that was used for the elicitation of speech is a story from Aesop's fables, "The North Wind and the Sun," which consists of phonetically rich sentences with a certain degree of length. The evaluation of speech is one of the essential factors in the L2 speech database. It is fairly difficult for raters to evaluate learners' speech when faced with relatively short utterances. Lastly, but not least, the raters who evaluated the speech in the database are non-native English speakers. The eligibility of raters is one of the issues in L2 performance assessment; the experience and knowledge of raters can affect their rating behavior, for example, their rating severity and consistency. The raters in this database were experienced Japanese language teachers who were familiar with Asian English learners and the context of learning English in Asia. Furthermore, the raters in this database received rater training, and the evaluations given by the raters were examined based on MFRA; unreliable raters and evaluation items were excluded to obtain reliable evaluation scores.

8 Construction, implementation, and evaluation of automatic second language speech evaluation system

8.1 Introduction

This chapter introduces the construction, implementation, and evaluation of an automatic L2 speech evaluation system. 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," which consists of 113 words and five sentences. Furthermore, to assist examinees to understand the feedback, the evaluations given to examinees are categorical scores: A, B, and C. Therefore, the estimated rankings are calculated based on Neural Test Theory (NTT) instead of their estimated ability in Multifaceted Rasch Analysis (MFRA). Although MFRA is a useful technique for test developers to analyze items and raters, it is inappropriate in this case because examinees' ability is estimated and calculated on a logit scale. As Shojima (2008) mentioned, a test as a tool for measurement of individual ability is not a high resolution tool. We manage to put examinees on an ordinal scale, but if we place them on an interval scale, it does not mean much for non-mathematical learners. For example, we cannot tell the difference in ability between an examinee with a test score of 82 and one with an 80 using a range from 0 to 100. What we can do by using a test is group examinees into several levels. Furthermore, it is very difficult to give feedback to examinees that are placed on an interval scale. The automatic speech evaluation system to be constructed in this study aims to give examinees feedback according to the level that the system estimates. Hence, in this study, learners' ability is calculated, based on NTT (Shojima, 2008).

This chapter is organized in the following manner. First, the calculation process of the prediction formula is shown. Second, the scoring procedure is described. Third, the

structure of the system and the test-taking procedure are depicted. Finally, the evaluation of the system is discussed.

8.2 Confirmatory analysis[15]

Based on the results in the correlation studies reported in Chapter 6, the 101 read-aloud speeches in the Asian English speech database, which is described in Chapter 7, were analyzed. The learners' language backgrounds were Japanese, Korean, Chinese, Filipino, Thai, Khmer, Malay, Vietnamese, Cambodian, and Indonesian. They were either undergraduate or graduate students. Their average age was 23.46 years old with a SD of 4.42, and their average time spent studying English was 11.88 years with a SD of 5.41. The relationship between speech characteristics and evaluation scores was examined using multiple regression analysis (stepwise method).

The examinees' abilities were determined through the analysis presented in 6.3 and estimated by MFRA. These abilities are used as the criterion variable. The predictor variables were 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. 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 was .69. Figure 8.1 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.

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

[15] This section first appeared as 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.

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[16], however, we conclude that by using the learners' speech characteristics obtained in the previous analyses, we are able to predict reliable evaluation scores in the automatic L2 speech evaluation system.

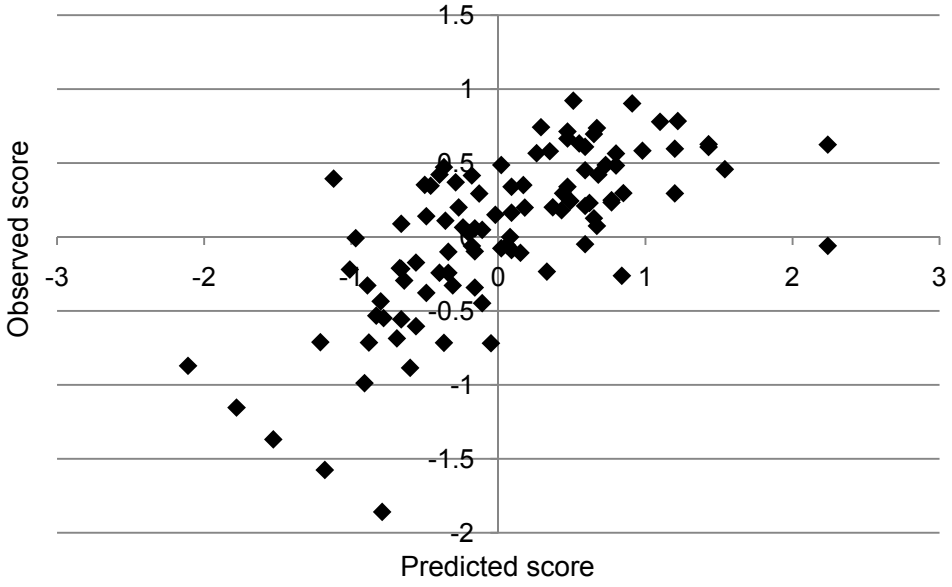

Figure 8.1 The Observed and Predicted Score

8.3 Level estimation based on Neural Test Theory

The evaluation scores of the speech in the Asian English speech database were re-analyzed, based on NTT to estimate the examinees' levels. The proposed automatic speech evaluation 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 Common European Framework of Reference (CEFR): basic users, independent users,

[16] The statistic indicating the proportion of variance in one variable that is predicted by the other, the square of the correlation coefficient.

and proficient users. In this analysis, the levels were set up to three, and the fit of the data to the model was examined.

The examinees were divided into three groups: thirty-six proficient users, thirty-one independent users, and forty-four basic users. Table 8.1 shows the test fit indices in NTT. The indices shown below all indicate the data's goodness-of-fit to the model in NTT (See 2.7 for the details of NTT).

Table 8.1 Test fit indices in NTT

Index	Value
χ^2_{156}	237.655
CFI	0.994
RMSEA	0.021

8.4 Scoring procedure

In Figure 8.2, the ranked speech data of the Asian English speech database are identified by using the values of pruned syllables per second and the average ratio of weak syllables to strong syllables. The x-axis indicates the value of the average ratio of weak syllables to strong syllables, and the y-axis indicates the values of pruned syllables per second. The values of the average ratio of weak syllables to strong syllables are inverted (the plotted values are 1 minus the original values) for a clearer picture. Although some outliers were found, and there is an area that is occupied by all three ranks, the areas that are appropriate to each rank can be defined to some extent. The averages of the two values were calculated in each category and plotted in Figure 8.3. The bigger indicators are the averages (prototypes) of the two values in each category.

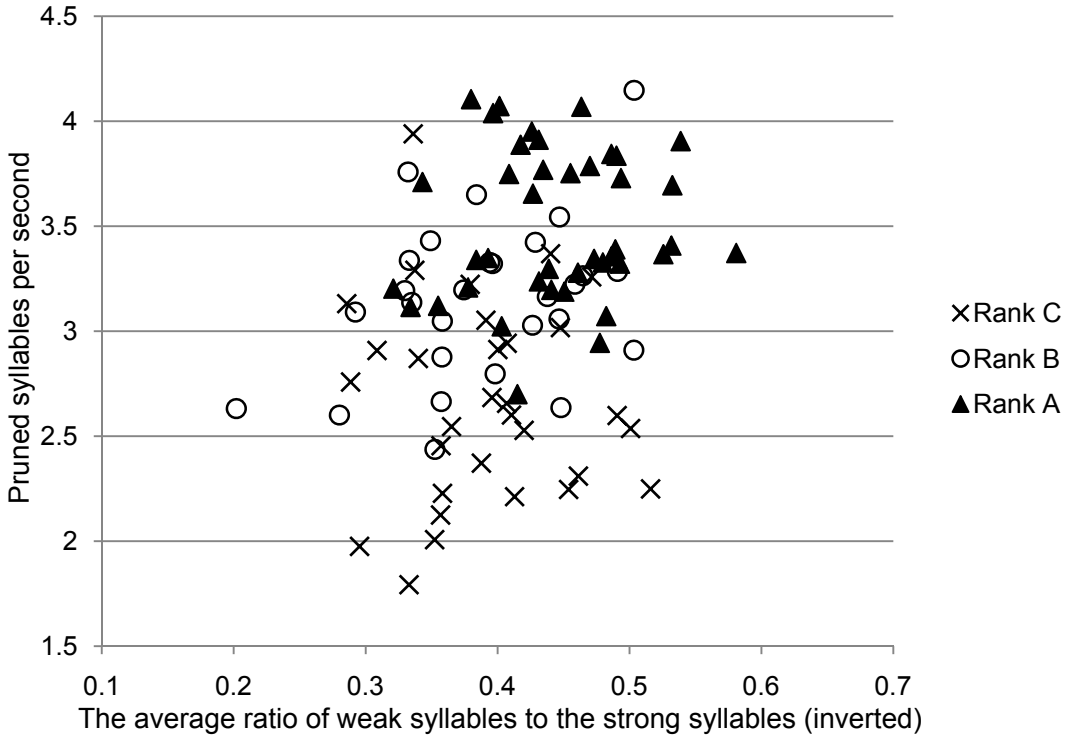

Figure 8.2 Scatter Graph for the Values of Pruned Syllables per Second and the Average Ratio of Weak Syllables to Strong Syllables in each Category

In the automatic evaluation system, a new examinee's category is determined based on the Euclidean distances to the prototypes in each category. The distances to each prototype are calculated using the equation below:

$$D(x, p) = \sqrt{(p_1 - x)^2 + (p_2 - y)^2} \tag{8.1}$$

where p_1 is the average of pruned syllables per second in a category, x is a new examinee's pruned syllables per second, p_2 is the average of the ratio of weak syllables to strong syllables, and y, the new examinee's average ratio of weak syllable to strong syllables. By comparing the distance of the new examinee's values to each prototype (Rank A, B, and C), the examinee is assigned to the category closest to their value.

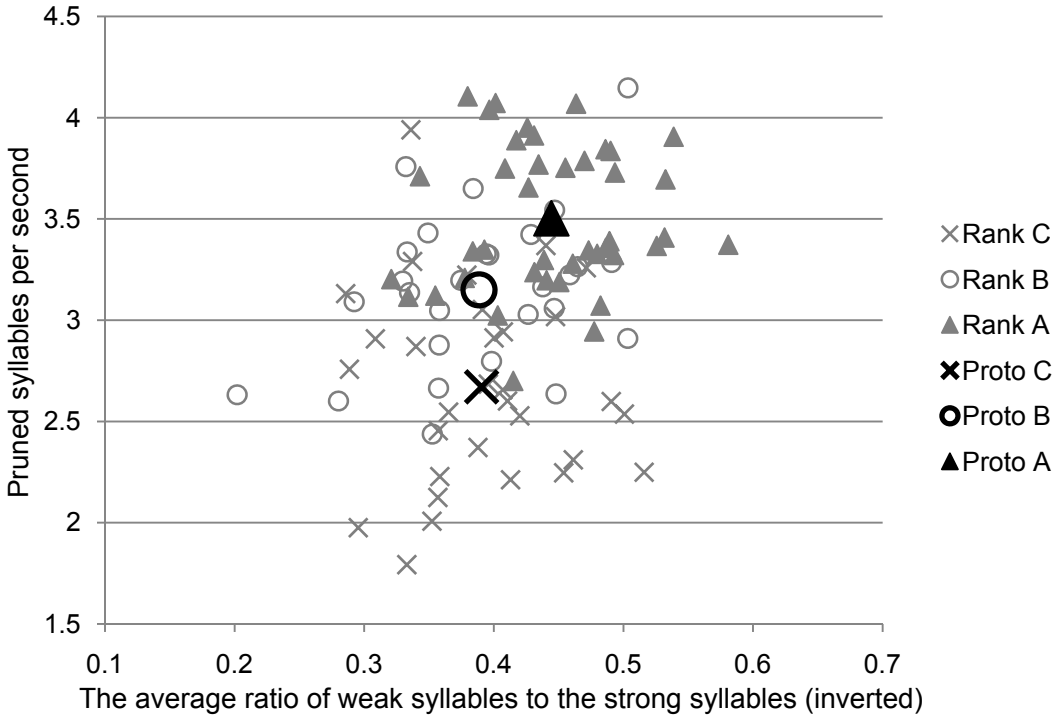

Figure 8.3 The Averages of the Values of Pruned Syllables per Second and the Average Ratio of Weak Syllables to Strong Syllables in each Category

8.5 Structure of the system

The automatic L2 speech evaluation 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 8.4 depicts the automatic evaluation procedure.

The system records an examinee's speech using the Java applet, JavaSonics ListenUp (Mobileer, Inc, 2008), and this recorded speech is transferred to the sever computer and stored. Then, the speech is converted to the 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 (e.g., Internet Explorer, FireFox, or Google Chrome).

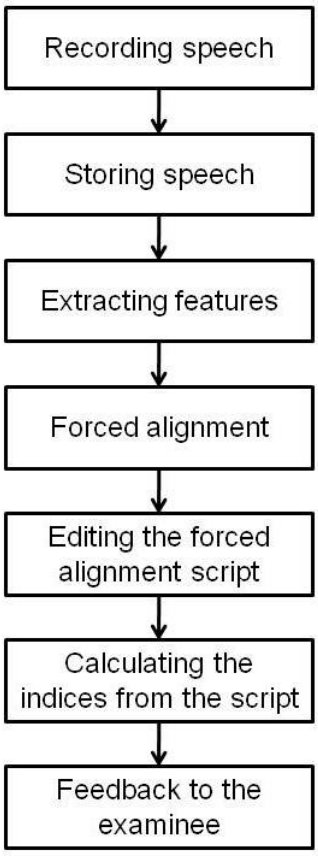

Figure 8.4 Procedure of Automatic evaluation

Figure 8.5 shows the structure of the evaluation system. The recording procedure is performed using the Java applet stored in the folder `codebase`, and the transfer and retention of the speech file are controlled by `upload_x.cgi`. The examinees record their speech sentence by sentence in a process that repeats five times. In each process, `upload_x.cgi` calls HTK to convert the speech file and phone-align the speech by using the files stored in the `hmm` folder. The HMM in this directory was trained with native speakers and learners' speech: See 7.3.

Figure 8.5 The Structure of the Evaluation Website

The converted speech file is stored in the directory `mfcc`, and the output file of the phone alignment is stored in the directory `out`. The output file of the phone-aligned speech is edited by `upload_x.cgi` and stored in the directory `lab`. Lastly, `eva.cgi` calculates the two indices of speech characteristics and the distances to the prototypes stored in the information in the edited output file in the directory `out`, and this produces the evaluation and feedback to the screen on an examinee's client computer. The directory `img` contains the

image file that was used in the webpage for examinees. The file `instruction.cgi` creates the instruction page, the file `recx.cgi` creates each recording page, and the file `testrec.cgi` controls the test recording in the instruction page. The files `data.dat` and `ques.dat` store the information that the examinees enter in the initial page. All of the scripts are listed in Appendix H.

8.6 Test-taking procedure

This section introduces the procedure of the automatic evaluation. Firstly, examinees access the evaluation website, enter their names, and answer a questionnaire. Figure 8.6 shows the initial page. 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 8.7 shows a screenshot of the instruction page. Figure 8.8 shows a screenshot of the recording page.

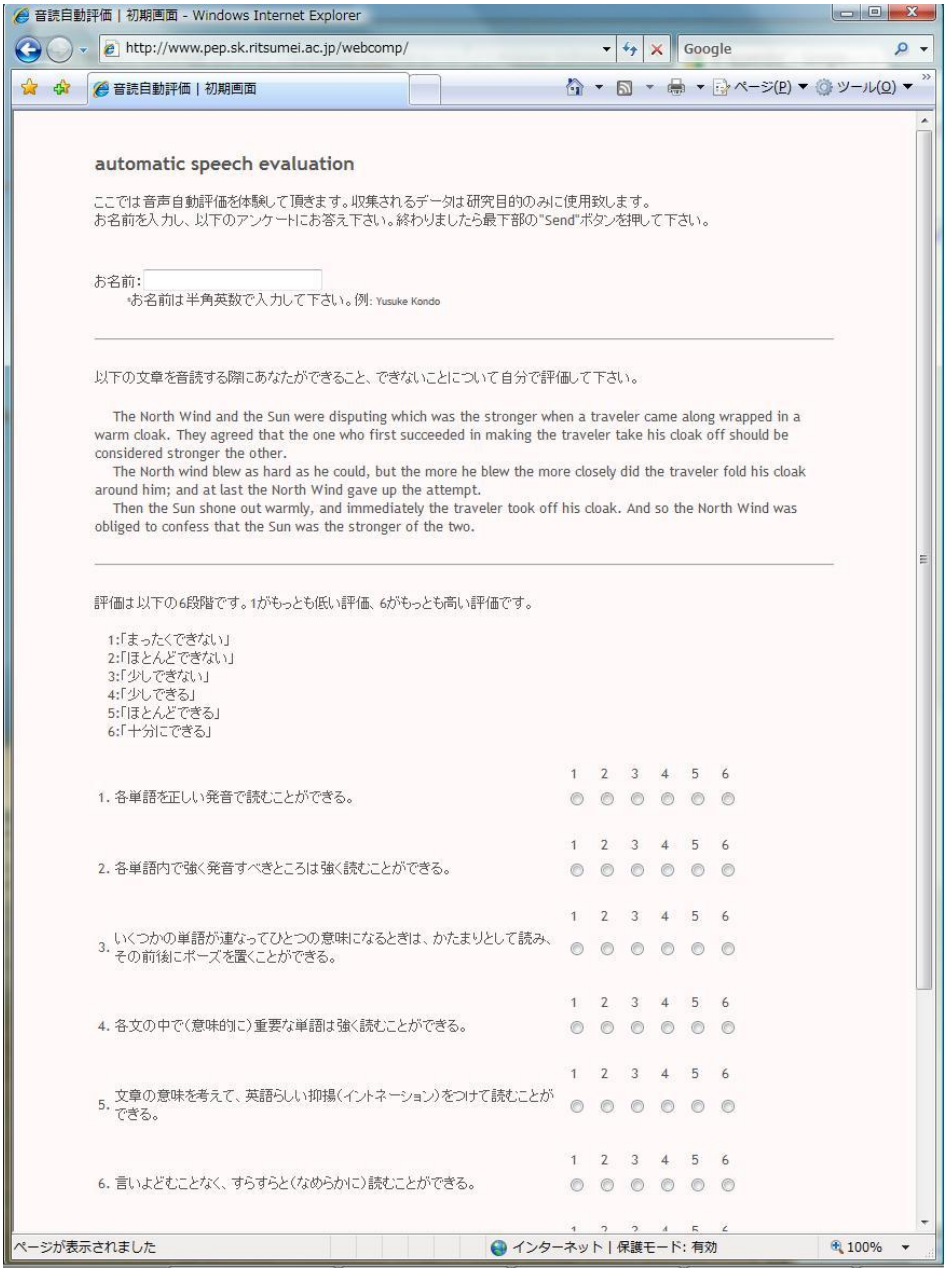

Figure 8.6 Questionnaire Page

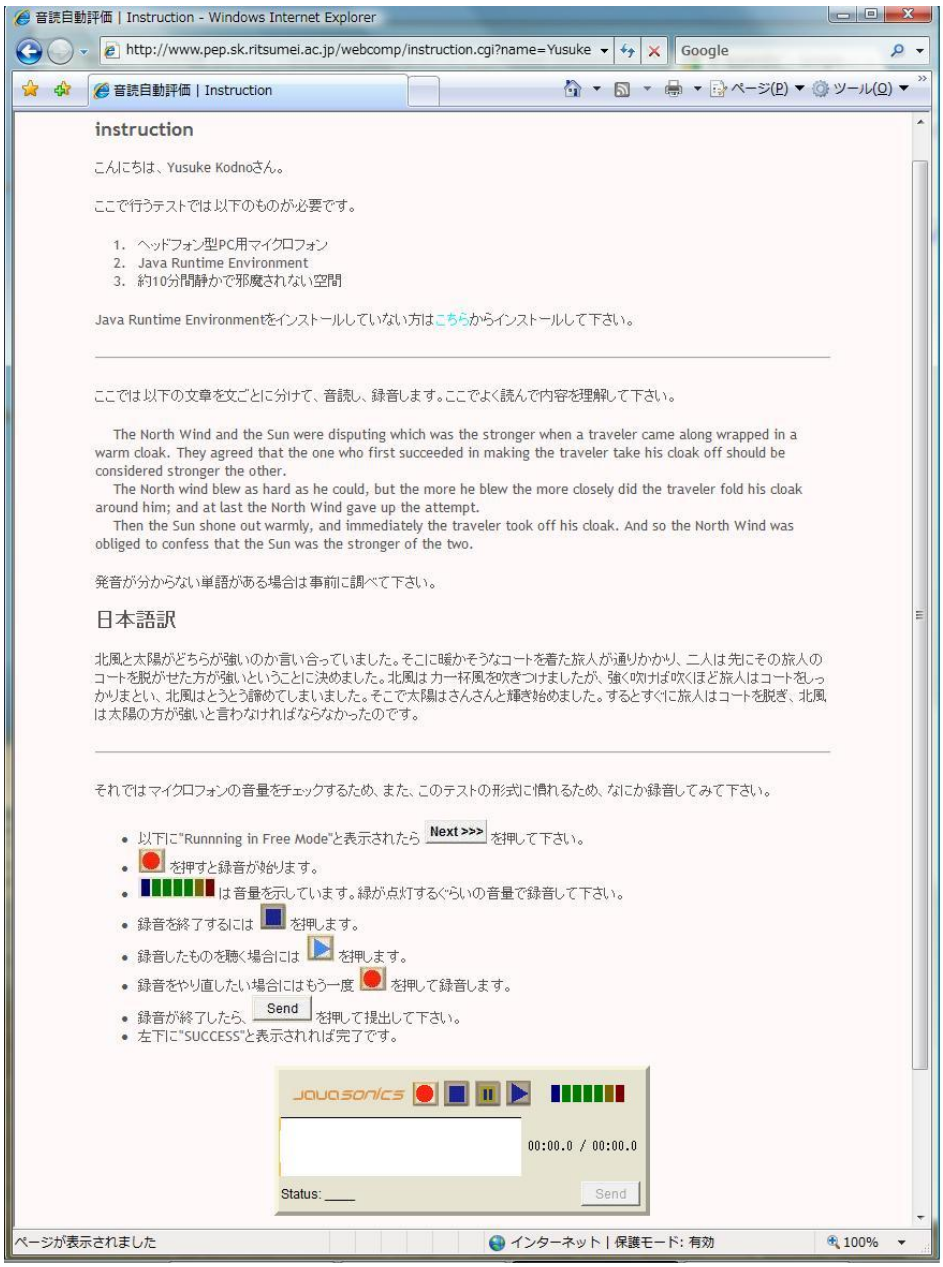

Figure 8.7 Instruction Page

Figure 8.8 Recording Page

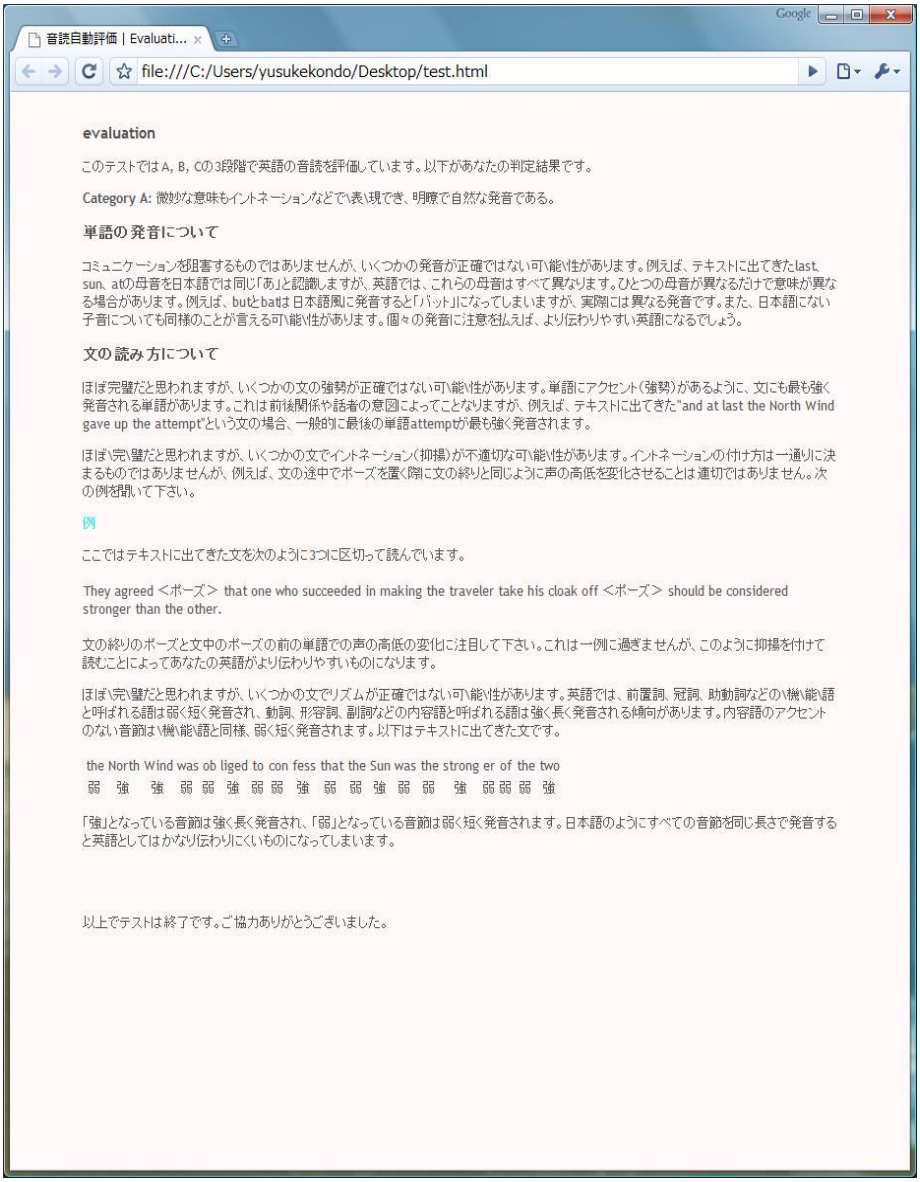

Figure 8.9 Evaluation Page

Lastly, after the examinees complete the recordings, they receive an evaluation of their speech and feedback according to their level, which the system estimates. Figure 8.9 shows an example of the evaluation page.

8.7 Evaluation of the system[17]

8.7.1 Introduction

Each speech data in Asian English speech database described in Chapter 7 contains the evaluation scores estimated by MFRA and phone-aligned utterance. Through the correlation studies reported in Chapter 6, two speech timing control characteristics, the indices of speech rate and rhythm, were found to be statistically significant predictors of the evaluation scores in read-aloud speech. In the automatic evaluation system proposed here, using these two predictors, examinees are categorized into three levels. In this section, firstly, as the preliminary stage for automatic scoring, using the speech data in Asian English speech database, methods for grouping the speech data are examined, and then, based on the results of the examination of the categorization, automatic scoring methods are tested, adopting new speech data of Japanese learners' of English in terms of the degree of agreement with human raters. The participants consisted of twenty one Japanese English learners and three raters. The raters were Japanese English language teachers who received rater training according to CEFR: see Chapter 3. The learners were Japanese undergraduate students. The participants evaluated their ability to read aloud by using evaluation items along a 6-point Likert scale. The reliability of the scores by the automatic evaluation system is examined by investigating the relationship between the evaluation scores and the self-evaluation scores by the examinees.

8.7.2 Examination of the methods for grouping the speech data

The proposed automatic speech evaluation system is a system that is meant to predict the

[17] A part of this section first appeared in 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 and Ueda, N., Mikami, A., Nakano, M., Kondo, Y., Tsutsui, E. (2010, Spe.). ICT katsuyou jugyou to jyugyou hyouka. [ICT-based English Courses and Assessment Issues]. Paper preented at JACET 48th Convention, Hokkaido, Japan.

evaluations given by human raters. Considering the reliability of human rating and the accuracy of the 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. Two methods for level estimation are examined: the categorizations based on NTT and Classical Test Theory (CTT).

The evaluation scores of the speech in the Asian English speech database were re-analyzed based on NTT to estimate the examinees' levels. In this analysis, the levels are set up to three, and the fit of the data to the model is examined: see 8.3. Figure 8.10 shows the three-ranked speech data in Asian English speech database which are identified with the two predictor variables, pruned syllables per second and the average ratio of weak syllables to the strong syllables.

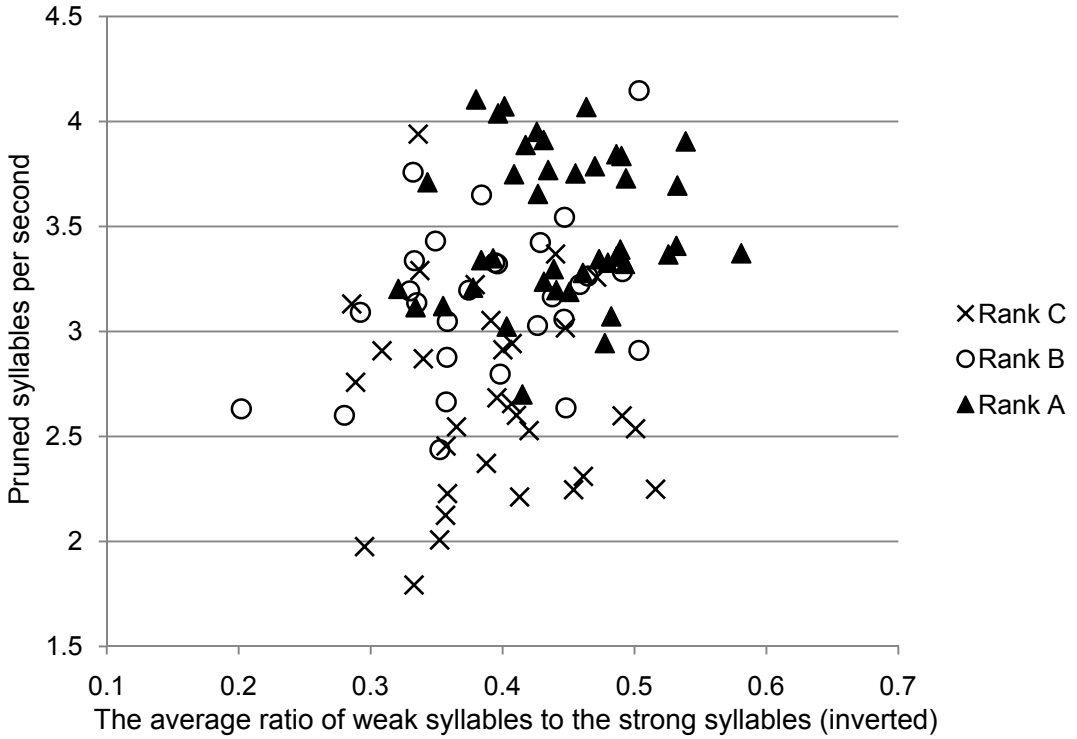

Figure 8.10 The Three-ranked Speech Data Based on NTT

To examine the accuracy of the discrimination by this method, the degree of agreement were

calculated. Firstly, the averages of the two predictor variables were calculated in the speech data of the three levels estimated by NTT: this average point is called “prototype”, and the distances from each data to the prototype in each level were measured. All the data were categorized into the level whose prototype was the nearest to the data. Then, the degree of the agreement between the levels estimated by the distance to the prototype and the levels estimated based on NTT. In this analysis, 64.35 per cent of the speech data was judged to be the same levels in these two methods.

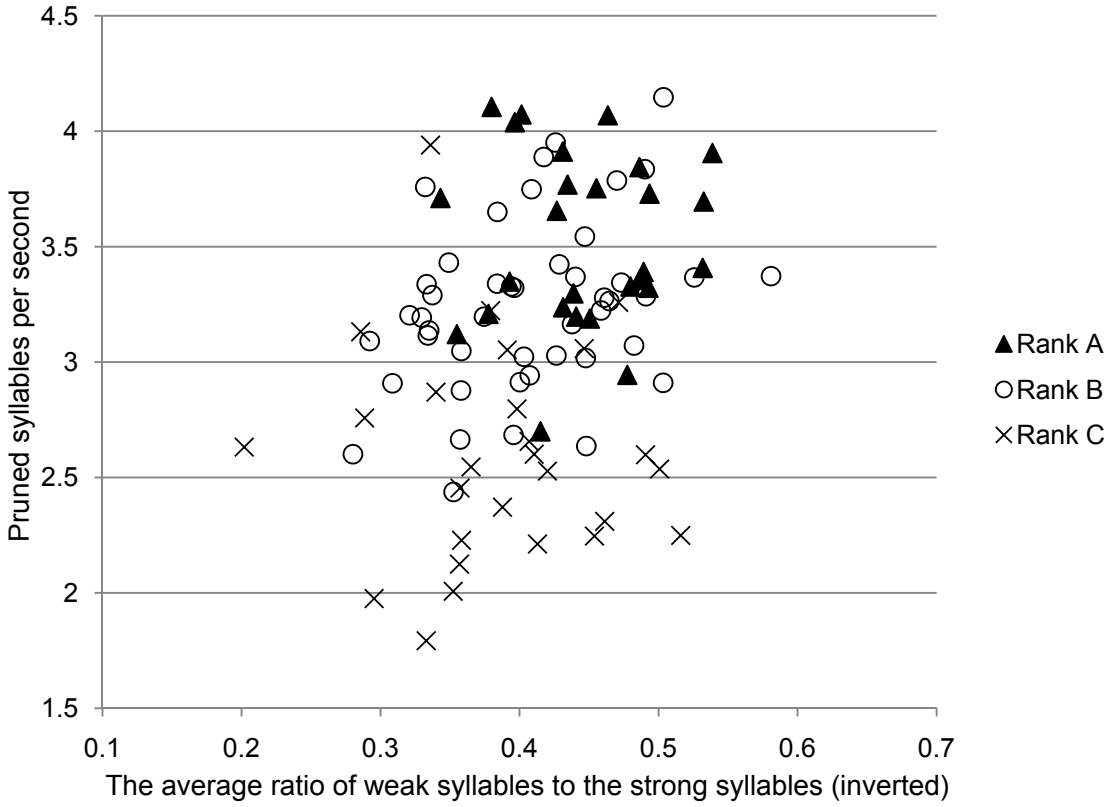

Figure 8.11 The Three-ranked Speech Data Based on CTT

The speech data in Asian English speech database contain the evaluation scores estimated based on MFRA: See 6.3. Based on the scores, the speech data were divided into three levels: 27 per cent of upper-level, 46 per cent of middle-level, and 27 per cent of lower-level. This method of grouping is used to analyze test items in CTT (Otomo, 1996). Figure 8.11

shows the three-ranked speech data which are identified with the two predictor variables, the indices of speech rate and rhythm. The same procedure was adopted to examine the accuracy of the discrimination by this method. The degree of the agreement between the levels estimated by the distance to the prototype and the levels estimated based on CTT. In this analysis, 61.38 per cent of the speech data was judged to the same levels in these two methods. In the present data, slightly better correct discrimination rate was obtained in the level estimation based on NTT.

8.7.3 Examination of scoring methods

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 (Shakhnarovich, Darrell, and Indyk, 2006) 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, and the correlation of the scores with the self-evaluation by the examinees.

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 characteristic beforehand, and a new data is grouped into the category according to its amount of characteristic. In NN method, prototypes are decided by calculating the averages of amount of characteristic 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 are measured, and the distance from the new examinee to the prototypes of three levels are calculated. Then, 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 multiple regression, based on the evaluation scores of the speech data in Asian English speech data base, which were estimated by MFRA: See 6.3, 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 (See 8.2). 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. Then, the relationships were examined among the self-evaluation by the examinees and the

evaluation scores given by the human raters and the proposed system.

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 automatic evaluation system. The evaluation scores by the human raters and the three scoring methods are depicted in Appendix I.

Fleiss' kappa (Fleiss, 1971) 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 (Gwet, 2001). The interpretation of this index is somewhat controversial, because it depends on the number of raters, categories, and examinees. The Fleiss' (1981) 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" (Fleiss, ibid: 218). Table 8.2 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 8.3 shows the kappa among the three human raters and the NN method. The indices were calculated four times. Each time, one of the raters was excluded. By comparing these indices, the rater who lowers the degree of agreement can be detected. For example, the kappa in the second row indicates the rater agreement among Raters 1 and 2 and the NN method, excluding Rater 3. The kappa in the lowest row indicates the rater agreement among all the raters: Raters 1, 2, 3, and the NN method.

Table 8.2 Fleiss' kappa among the human raters and the three scoring methods

Method	κ
NN method	.66
k-NN method	.42
Multiple regression	.49

Table 8.3 Fleiss' kappa among the raters

Raters	κ
Rater 1, 2, and NN method	.70
Rater 1, 3 and NN method	.60
Rater 2, 3, and NN method	.60
Rater 1, 2, and 3	.75
ALL	.66

Table 8.4 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 8.5 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 according to Formula (8.2) (Shiotani and Asano, 1967:195). Formula (8.3) re-transformed the values into the correlation coefficient. The average of the inter-rater reliability in this evaluation is .79.

Table 8.4 Correlation coefficients among the raters and the three scoring methods

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

Table 8.5 The correlation coefficients between the human raters and the system

	NN method	Rater 1	Rater 2	Rater 3
NN method	1	.81	.69	.58
Rater 1		1	.83	.80
Rater 2			1	.89
Rater 3				1

$$z = \frac{1}{2}\log e(1+r_{xy}) - \frac{1}{2}\log e(1-r_{xy}) \tag{8.2}$$

$$r_{xy} = (\exp\ 2z - 1)/(\exp\ 2z + 1) \tag{8.3}$$

8.7.4 Self-evaluation score

Before the participants took the test through this system, they evaluated their own ability to read "The North Wind and the Sun" aloud. The items by which the participants evaluated themselves are listed in Table 8.6. These items were created based on the evaluation items in 6.2.2. Some items were excluded and others were altered to assist the participants' understanding. The participants evaluated themselves by rating these eight items along a 6-point Likert scale. The theoretical range of this scale is eight to forty eight.

In this experiment, although this score is a self-evaluation score of the ability to read out

loud, the relationship between the score and the evaluation given by the system and the human raters was examined. The items were presented to the examinees in Japanese. The Japanese versions of the items are listed in Appendix J.

Table 8.6 Items for self evaluation in reading-out

1. I can read out every word in the passage accurately with good pronunciation.
2. I can read out every word in the passage accurately with good rhythm.
3. I can make pauses in the passage based on the meaning of the passage.
4. I can read out the passage accurately while putting the stress on the important words.
5. I can read out the passage accurately with good intonation.
6. I can read out the passage fluently.
7. I can understand the meaning of the passage while reading it.
8. I can understand the grammar of the passage while reading it.

The descriptive statistics of the items in the self-evaluation score are shown in Table 8.7. In the first and the second column, the means and the standard deviations of each item are listed, respectively. In the third column, the correlation coefficient of the score of the item in question with the sum of the remaining items is listed. In the forth column, the Cronbach's alpha is listed for all the items if the item was excluded. Although this scale consists of only eight items and a relatively small number of participants answered these items, there was no item that lowered the reliability of the scale. This scale is considered to measure one construct: self-evaluation of the ability of reading aloud in L2. Table 8.8 shows the average scores of the examinees categorized into the three levels by the human raters and the three scoring methods. The average scores are also shown in Figure 8.12. The same tendency

was found among the average scores in each group categorized by the human raters and the three scoring methods.

Table 8.7 Descriptive statistics of the items in self-evaluation score in reading aloud

	M	*SD*	*R*	*α* if item excluded
Item 1	4.15	1.18	.89	.96
Item 2	3.65	1.42	.89	.96
Item 3	3.50	1.60	.88	.96
Item 4	3.55	1.50	.91	.96
Item 5	3.40	1.60	.95	.96
Item 6	3.55	1.47	.94	.96
Item 7	3.60	1.60	.80	.97
Item 8	3.55	1.43	.74	.97

As shown in Figure 8.12, both in the evaluation of the human raters and of the three scoring methods, there were almost no difference in the average scores between B level and C level. These results might be attributed to the fact that, as shown in Figure 8.2, the large multi-occupied area by B and C level are found in the data which were used to decide the three levels.

Table 8.8 The average scores of the examinees categorized into the three levels

	A	B	C
Rater 1	36.63	23.00	24.25
Rater 2	39.67	25.25	23.17
Rater 3	45.75	24.44	25.14
NN method	37.00	25.40	22.75
k-NN method	41.67	31.71	23.45
Multiple regression	32.14	21.75	22.67

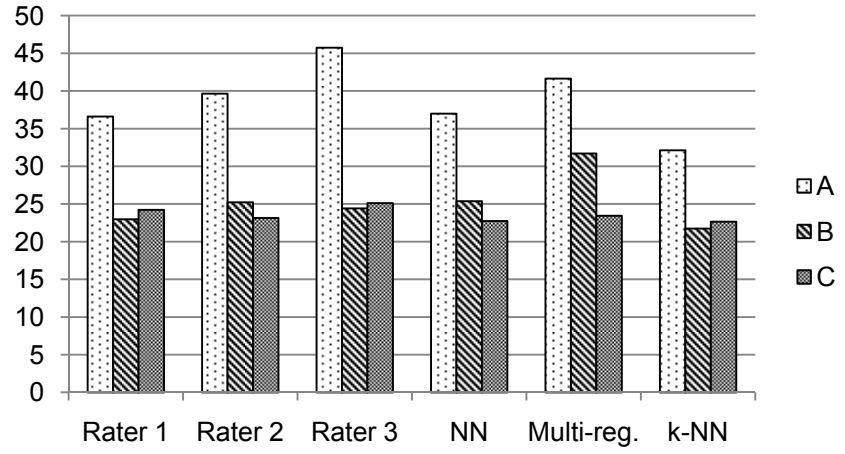

Figure 8.12 The Average of the Self-evaluation Scores of the Examinees

8.8 Summary and discussion

This chapter 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 had been given evaluation scores by trained human raters. The evaluation in the system was 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 from two points of view: the degree of agreement among the system and human raters, and the relationship between the evaluations and the self-evaluation scores in the read-aloud speech. First, 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): See Table 8.3. Furthermore, the average of the correlation coefficients among the human raters and the system was fairly high (.79). Then, the average scores of the self-evaluation by the examinees in the read-aloud speech were categorized according to the evaluation by the system and the human raters. Similar tendencies were found in the evaluations by both human raters and system. Judging from the results of these two 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 now 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 L2 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-aloud 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. Furthermore, the average of the correlation coefficients among the human raters and the system is .79. 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.

In the second experiment, the relationship between the self-evaluation score and the levels given by the system and the human raters was examined. The system denoted the same tendency of the raters: the average score in level A is the highest, and there is little difference in the average score between levels B and C. As Figure 8.2 indicates, there is an area that is occupied by both levels B and C, and the area of level A is relatively independent from the other levels. It seems that this distribution reflects the tendency of the average scores of the examinees that were categorized by the system and the human raters.

9 Conclusion

9.1 Summary and conclusion

The aim of this study was to construct the automatic second language (L2) speech evaluation system which predicted the evaluation scores given by experienced language teachers who are L2 users of English. Common European Framework of Reference (CEFR) and European Language Portfolio (ELP) were adopted as the criteria of L2 performance assessment in this study. In Chapter 3, the applicability of the levels in CEFR and the "Can-do" statements in ELP were examined in the context of language learning in Japan. 2619 Japanese university students self-evaluated their speaking ability, and their teachers evaluated their 982 students with reference to the "Can-do" statements in ELP. Their evaluation was analyzed based on Item Response Theory (IRT), two parameter logistic model. The results indicated that the six levels in CEFR were clearly differentiated both in the self-evaluation by the students and the evaluation by the teachers, though discrepancy was found between the teachers and the students in some of the "Can-do" statements. However, the discrepancy was related to the translation problem and the cultural differences between European countries and Japan, not related to the speaking ability itself. Based on the results, we can conclude that the levels in CEFR and the "Can-do" statements in ELP are applicable to the context of language learning in Japan, a country where English is learned as a foreign language.

Applying the six levels in CEFR, rater training was conducted, and the training effects were examined in Chapter 4. Experienced language teachers received the rater training and evaluated L2 speech performance. The evaluations were analyzed based on Generalizability Theory (G-Theory) and Multifaceted Rasch Analysis (MFRA) before and after the training. Although the analysis based on MFRA revealed no difference in rater characteristics before and after the training, the analysis based on G-Theory indicated that the effects of the training

was the reduction of the source of error related to the evaluation items. The variance related to the items was reduced to about one sixth after the training. This is because the raters understood the content of the evaluation items better through the training. The raters seemed to be familiarized with the rating procedure through the training. This study demonstrated the importance of rater training in L2 speech performance.

Through the studies reported in Chapters 5 and 6, the evaluations of the spontaneous and the read-aloud speeches by the trained raters were analyzed based on MFRA, and several items were excluded to obtain reliable evaluation scores. While three items were excluded, based on their fit statistics in the evaluation of the spontaneous speech, no item was excluded in the evaluation of the read-aloud speech.

In the examination of the predictability of the evaluation scores by the speech characteristics in the evaluation of the spontaneous speech reported in Chapter 5, high correlations were not found between the evaluation scores and the indices of lexical richness and syntactic accuracy and complexity. However, the speech-timing control characteristics (e.g. word per minute and number of filled pause) were found to be statistically significant predictors of the evaluation scores. The results lead us to conclude that the evaluation by human raters is more likely to be affected by the timing-control characteristics than by lexical richness and syntactic features in the present data. Based on the results, the predictability of the evaluation scores by the speech characteristics was examined in the read-aloud speech, which was reported in Chapter 6. The target features were timing-control characteristics, pause control, vowel discrimination, vowel reduction, loudness, pitch, and pronunciation error. Several characteristics were found to fairly correlate with the evaluation scores in the read-aloud speech, but only two indices were verified as the statistically significant predictors of the evaluation scores: pruned syllable per second and the average ratio of the weak syllables to the strong syllables.

Based on the results of the correlation studies in the read-aloud speech, the automatic L2 speech evaluation system was constructed. The speech data used to train acoustic model in a speech recognizer were 101 Asian learners of English whose details were described in Chapter 7. In the automatic L2 speech evaluation system, the speech data were grouped into three levels: A, B, and C according to the evaluation scores based on Neural Test Theory (NTT). NTT is a test theory which assumes examinees' ability to be scaled as ordinal. This was a method congruous with the present evaluation, because the evaluations by the raters were based on the levels in CEFR, and the feedback given by the automatic evaluation system would be categorical in nature. In the system, the read-aloud speech data is processed by two speech characteristics, the pruned syllable per second and the average ratio of the weak syllables to the strong syllables, which were verified as the statistically significant predictors of the evaluation scores in the correlation studies, and the averages (prototypes) of the two speech characteristics in each rank of the speech data were calculated. The two speech characteristics of new examinees were detected by the speech recognizer, and the scores of new examinees were determined on the basis of the Euclidean distances to the prototype in each category. Comparing the three distances of the new examinees' values to each prototype ranks A, B, and C, the examinee is given the category of the nearest distance. The system operates via the internet, which an examinee takes the test in any place where the internet is available.

The evaluation of the system was done from two perspectives: the degree of agreement with human raters and the correlation with the self-evaluation in the read-aloud speech by examinees. The degrees of the agreement of the system with the human raters were found to be fairly high, and the similar tendency was found in the relationship between the self-evaluation scores and the evaluations by the system. The results imply that the system can work as one of the human raters in the evaluation of L2 read-aloud speech.

The defining characteristic of the study is that the automatic L2 speech evaluation system was constructed by adopting approaches and technology in Speech Science, Educational Measurement, and Applied Linguistics. A few research projects have been conducted, collaborating on the development of automatic speech evaluation system (e.g. Xi, et al., 2008; Bernstein, 1999). Until recently, speech scientists attempted to construct an automatic L2 speech evaluation system and to improve the accuracy of the scores produced by the system. Researchers and practitioners in Applied Linguistics did not pay attention to measurement models in their tests. They did not apply technology of Speech Science in the measurement of learner language. However, to construct an automatic L2 speech evaluation system to produce a reliable score, we need to apply the approaches and the insights of these disciplines.

The present study demonstrated the applicability of the criterion for the evaluations (CEFR) to the context of English language learning in Japan and examined the reliability and the consistency of raters and evaluation items, based on the test theories, G-Theory and MFRA. In the evaluation of read-aloud speech, furthermore, a text with a certain length was adopted for raters to catch the learners' speech characteristics. L2 users were employed as raters, namely, Japanese teachers of English with experience, considering the present situation of English language learning. The correlation between the evaluation scores delivered through this evaluation and the speech characteristics were examined, and the automatic evaluation system based on the results of the correlation studies were constructed. The results of the evaluation of the system indicate the system is capable of delivering reliable scores in L2 assessment of read-aloud speech.

9.2 Limitations of the study and directions for future research

The present study contains certain limitations which need to be taken into account when considering its findings and contributions.

CEFR was used as the criteria in the human rating of this study, but there had not been sufficient discussion over the selection of criteria to evaluate the spontaneous and the read-aloud speech by Asian learners of English. Although several criteria are now available for language teaching, learning, and assessment, there is no criterion that provides resources to relate the objective criterion to an assessment. As for CEFR, however, a variety of documentations is available: the descriptors in the levels, the "Can-do" statements, the manual for relating an assessment to CEFR, and the supplement of the manual, which describes statistic methodology to analyze evaluations. These documentations are essential for raters to understand the procedure of rating and learners' levels set in the criterion. Furthermore, the video which describes examples of learners with the six levels in CEFR is available. The examples are helpful for raters to understand the learners of the six levels in CEFR.

In the evaluations of the present study, Japanese teachers of English evaluated the learners with a variety of language backgrounds through Asia: Japanese, Chinese, Korean, Thai, Vietnamese, and so on. If a rater shares his/her first language with examinees, a difference may appear in the evaluation. As reported in 2.3, Kim (2009) investigated the differences in L2 evaluation of oral proficiency of English between native speakers and non-native speakers of English. She found no difference between them in terms of self-consistency of raters'. In the present study, a clear criterion was adopted in the rater training, and the evaluations were analyzed, based on statistic models to detect unreliable raters and evaluation items. We may conclude that reliable evaluation scores were produced through the procedure.

Another problem in the evaluation is found in the selection of evaluation items and task type. In the present study, the evaluation items both in the spontaneous and the read-aloud speech were selected based on the discussion among the raters, and were scrutinized in the analyses by MFRA. Since the analyses were based on Item Response Theory, and misfit items were excluded through the analysis, the reliability of the evaluation was statistically

demonstrated. There was no discussion on the constructs measured in the spontaneous and the read-aloud speech. However, as North and Schneider (1998: 242) mentioned, there is no language proficiency models that are empirically and theoretically valid. Therefore we can only select evaluation items based on the raters' experience and statistics. In the present evaluations, experienced language teachers discussed the contents of the evaluation items, and the items were scrutinized, based on MFRA. The aim of the proposed automatic evaluation system was to predict evaluation scores by experienced language teachers. As for this aim, we obtained the reliable evaluation scores through the human rating. In addition, the present study concentrated on the two elicitation task, spontaneous speech (self-introduction) and read-aloud speech in the examination of learners' speech and their evaluations. However, it is a well-known fact that learners' speech is potentially impacted by task type and its contents. Hence, further researches need to be conducted in order to develop a comprehensive system that scores L2 speaking ability.

A problem in the proposed automatic evaluation system is its scoring method. In the system, an examinee's score is determined based on the Euclidean distances to the averages in the categories (A, B, and C) of the 101 speech data of Asian learners of English. This method is completely dependent on the existing speech data. The averages are subject to characteristics of the existing speech data. Although the speech data were collected from Asian learners of English with variety of first languages, there must be a possibility to obtain different results in the evaluation of the proposed system if data collection had been done with care of the variety of first language and proficiency. However, as in the results of the evaluation of the system reported in 8.7 shows, the scores produced by the system are substantially correlated with the scores by the human raters. It leads us to conclude that we can obtain reliable evaluation scores in the read-aloud speech by using the proposed system.

References

Akiyama, T. (2001). The application of G-theory and IRT in the analysis of data from speaking tests administered in a classroom context. *Melbourne papers in language testing*, 10 (1).

ALC Press (2006). SST: Standard Speaking Test. Retrieved May 30, 2010, from http://www.alc.co.jp/edusys/sst/english.html.

American Council for the Teaching of Foreign Languages. (1999). *ACTFL Proficiency Guidelines*. Retrieved March, 20, 2010, from http://www.sil.org/lingualinks/languagelearning/OtherResources/ACTFLProficiencyGuidelines/contents.htm.

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.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: CUP.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, 12 (2), 238-257.

Bernstein, J. (1999). PhonePass testing: Structure and construct. Menlo Park, CA: Ordinate.

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.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20 (1), 89-110.

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

Canagarajah, A. S. (1999). Interrogating the “native speaker fallacy”: non-linguistic roots,

non-pedagogical results in Braine, G. (ed.): *Non-native educators in English language teaching*. Mahwah, NJ: Lawrence Erlbaum Associates.

Centre for Canadian Language Benchmarks (2000). *Canadian language benchmarks 2000*. Retrieved March 20, 2010, from http://www.language.ca/pdfs/clb_adults.pdf.

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

_____. (2003). *Relating language examinations to the Common European Framework of Reference for languages: learning, teaching, assessment* (CEF). Manual: Preliminary Pilot Version. DGIV/EDU/LANG 2003, 5. Strasbourg: Language policy division.

_____. (2005). *Reference supplement to the preliminary version of the manual for relating language examinations to the Common European Framework of Reference for languages: learning, teaching, assessment* (CEF). Manual: Preliminary Pilot Version. DGIV/EDU/LANG 2003, 13. Strasbourg: Language policy division.

Crick, J. E., & Brennan, R. L. (1984). GENOVA: A general purpose analysis of variance system. Version 2.2 [Computer software]. Iowa: American College Testing Program.

Crystal, D. (2003). *The Cambridge encyclopedia of the English language*. Cambridge: CUP.

Cucchiarini, C., Strik, H., & Boves, L. (2000a). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication, 30*, 109-119.

_____. (2000b). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of Acoustical Society of America*, 107, (2). 989-999.

_____. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of Acoustic Society of America*, 111 (6), 2862-2873.

Daigaku Eigo Kyoiku Gakkai Kihon Go Kaicho Iin kai. (2003). *Daigaku Eigo Kyoiku Gakkai Kihongo List*. [JACET list of 8000 basic words]. JACET: Tokyo.

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.

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.

de Wet, F., Van der Walt, C., & Niesler, T. R. (2009). Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication, 51*, 864-874.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahawah: Lawrence Erlbaum Associates, Inc.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, 51, 832-844.

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.

Flege, J. (1987). A critical period for learning to pronounce foreign languages? *Applied Linguistics*, 8, 162-177.

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

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. D., Dahlgren, N. L., & Zue, V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus.

Philadelphia: Linguistic Data Consortium.

Guion, S.G. (2003). The vowel systems of Quichua-Spanish bilinguals: An investigation into age of acquisition effects on the mutual influence of the first and second languages. *Phonetica* 60, 98-128.

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

Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing*, 22, (3), 337-354.

Ikeda, H. (1994). Gendai test riron. [Contemporary test theory]. Tokyo: Asakura Shoten.

Iwahara, S. (1971). *Suikeigaku ni yoru shin kyouiku toukeihou*. [New statistical methods for education based on inferetial statistics]. Tokyo:Nihon Bunka Kagakusha.

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.

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, 25, 228-242.

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

Kim, Y-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, 26 (2), 187-217.

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

Kitagawa, A., & Kondo, Y. (2008). Reduction of vowels by Japanese learners of English. *Proceedings of 13th Conference of Pan-Pacific Association of Applied Linguistics*, 227-230.

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

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19, (1), 3-31.

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.

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.

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.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, 21 (1), 1-27.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, vol. 19. 3-31.

Kunnan, A. J. (1992). An investigation of a criterion- referenced test using G-theory, and factor and cluster analyses, *Language Testing*, 9 (1), 30-49.

Lee, J. F., & Musumeci, D. (1988). On hierachies of reading skills and text types. *Modern Language Journal*, 72, 173-187.

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

_____. (2006) Facets Rasch measurement [Computer software]. Chicago: Winsteps.com.

Little, D. (2002). *The European language portfolio: structure, origins, implementation and challenges*. Language Teaching, 35, 182-189.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, (12) 1. 54-71.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3, 4, 331-345.

McKay, S. L. (2002). *Teaching English as an international language*. Oxford: OUP.

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

Minematsu, N., Tomiyama, Y, Yoshimoto, K., Shimizu, K., Nakagawa, S., Dantsuji, M., & Makino, S. (2003). The construction of English read-speech database of Japanese and Americans for CALL [Eigo CALL kouchiku wa mokuteki tosita nihonjin oyobi beikokujin niyoru yomiage eigo onsei database no kouchiku]. *Journal of Japan Society for Educational Technology*, 27 (3), 259-272.

Mobiller, Inc. (2006). ListenUp SDK [Computer program]. Retrieved May 30, 2010, from http://www.javasonics.com/

Morrow, K. (2004). *Insights from the Common European Framework*. Oxford: OUP.

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

Myford, C. M. & Wolfe, E. W. (2004a). Detecting and measuring rater effects using many-face Rasch measurement: Part I. In Smith, Jr., E. V., & Smith, R. M. (eds). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.

460-517.

_____. (2004b). Detecting and measuring rater effects using many-face Rasch measurement: Part II. In Smith, Jr., E. V., & Smith, R. M. (eds). *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press. 518-574.

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*. 8 pages in CD-ROM Proceedings.

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.

Nakano, M., Kondo, Y., Tsutsui, E., & Owada, K. (2007, June). Daigaku eigo kyouiku ni okeru koutou happyou nouryoku no hyouka to sokutei. [Evaluation and Measurement of Second Language Speech in the University Context: Towards an Automatic Evaluation System]. Paper presented at 2007 Convention of the Japan Association of College English Teachers Kanto Chapter, Tokyo, Japan.

Neumeyer, L., Franco, H., Digalakis, V., Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication*, 30, 83-93.

Norcini, J. J., & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education*, 10 (1), 39-59.

North, B., & Hughes, G. (2003). CEF illustrative performance samples for relating language examinations to the CEF of languages: Learning, Teaching, Assessment (CEF) English (Swiss Adult Learners). Eurocentres.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, 14, (2), 217-263.

Osada, N. (2003). *The Effects of Silent Pauses on Listening Comprehension: A Case of Japanese Learners of English as a Foreign Language*. Unpublished doctoral dissertation. Waseda University: Tokyo.

Owada, K. (2005). Gakushusha corpus. [Learner corpus]. In Nakano, M. (ed.). Eigo kyoiku global design. [Global design in English language education]. Gakubunsha: Tokyo. 44-54.

Piske, T., Mackay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: review. *Journal of Phonetics*, 29, 191-215.

Raab, M., Gruhn, R., & Nöth, E. (2007). Non-Native Speech Databases. *IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU), Kyoto, December 2007, 413-418.

Roach, P. (2000). *English phonetics and phonology: A practical course*. Cambridge: CUP.

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

Salaberry, R. (2000). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing*, vol. 17. 289 - 310.

Schärer, R. (2000). *European language portfolio: final report on the pilot project*, Strasbourg: Council of Europe. http://culture.coe.int/portfolio.

Schärer, R. & Rapporteur, G. (2004). *A European Language Portfolio from piloting to implementation (2001-2004)*. Retrieved March 20, 2010, from http://www.coe.int/T/DG4/Portfolio/?L=E&M=/main_pages/documents.html

Schaden, S., & Jekosch, U. (2006). “Casselberveetovallarga” and other unpronounceable places: The CrossTown corpus. *Proceedings of LREC*, Genova, Italy,

993-998.

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

Shiotani, M., & Asano, C. (1967). *Tahenryokaisekiron* [Multivariate analysis theory]. Tokyo: Kyoritsu Shuppan.

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

_____. (2009) Exametrica 1.3 [Computer software]. (Retrieved December, 2009). http://www.rd.dnc.ac.jp/~shojima/exmk/jindex.htm

Sjölander, K., & Beskow, J. (2000). WaveSurfer - an open source speech tool. In Yuan, B., Huang, T., & Tang, X. (Eds.), *Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing*, 464-467. Beijing.

Takanashi, Y. (2009). *Data de yomu eigo kyouiku no joushiki*. [Understanding common knowledge of English language education by data]. Tokyo: Kenkyusha.

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

Toyoda, H. (1998). *Kyobunsan kouzou bunseki nyuumonhen*. [Covariance structural analysis: An Introduction]. Tokyo: Asakura Shoten.

_____. (2002). *Koumokuoutouriron nyuumonhen*. [Item Response Theory: An introduction]. Tokyo: Asakura Shoten.

_____. (2007). Kyobunsan konzou bunseki [Amos hen]. [Covariance structural analysis for Amos]. Tokyo: Tokyotosho.

Trim, J. L. M. (1978). *Some possible lines of development of an overall structure for a European unit credit scheme for foreign language learning by adults*. Council of Europe.

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.

Tsutsui, E., Kondo, Y., & Nakano, M. (2007) Nihonjin Eigo Gakushusha No Jissenteki Hatsuwa Noryoku Ni Kansuru Hyoko Kijun No Kento Common European Framework of References O Kiban Tosite. [An investigation on criterion for assessment of speaking ability of Japanese learners of English with reference to Common European Framework of References]. *Proceedings of the 5th Annual Conference of the Japan Association for Research on Testing*. 88-91.

Ueda, N., Mikami, A., Nakano, M., Kondo, Y., Tsutsui, E. (2010, Sep.). ICT katsuyou jugyou to jyugyou hyouka. [ICT-based English Courses and Assessment Issues]. Paper presented at JACET 48th Convention, Hokkaido, Japan.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15, 2. 263-287.

Widdowson, H. G. (2003). *Defining issues in English language teaching*. Oxford: OUP.

William, J. B., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, vol. 20, 89-110.

Wright, B. D., & Linacre, L. M. (1994). Reasonable mean-square fit values. *Rasch measurement: Transaction of the Rasch Measurement SIG*, 8, 370.

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.

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

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). *HTK book*. Cambridge University Engineering Department.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 24, 1-27.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication 51*, 883-895.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R.D. (2003). BILOG-MG 3 for Windows: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Appendix A: The original and translated versions of the “Can-do” statements in European Language Portfolio

Note. SI, SP, LS, and LQ stands for Spoken Interaction, Spoken Production, Language Strategy, and Language Quality, respectively.

SI01 I can introduce somebody and use basic greeting and leave-taking expressions.
他の人を紹介することや、基本的な挨拶や別れの際の表現を使うことができる。

SI02 I can ask and answer simple questions, initiate and respond to simple statements in areas of immediate need or on very familiar topics.
日常的なことやごく身近な話題ついて、簡単な質問や返答をしたり、簡単な発言や受け答えをすることができる。

SI03 I can make myself understood in a simple way but I am dependent on my partner being prepared to repeat more slowly and rephrase what I say and to help me to say what I want.
相手がゆっくりと話の内容を繰り返したり、自分の言ったことを言い直したりして手助けをしてくれれば、単純な方法で自分の意思を伝えることができる。

SI04 I can make simple purchases where pointing or other gestures can support what I say.
指差しやその他の身振りが意思疎通の手助けになる場合は、簡単な買い物をすることができる。

SI05 I can handle numbers, quantities, cost and time.
数や量、価格や時間を言うことができる。

SI06 I can ask people for things and give people things.

他の人にものを頼んだり、ものを与えたりすることができる。

SI07 I can ask people questions about where they live, people they know, things they have, etc. and answer such questions addressed to me provided they are articulated slowly and clearly.

他の人に住所、知人、持ち物などに関する質問をすることができる。また、ゆっくりとしたスピードではっきりと発音してもらえれば、自分に対するこうした質問に答えることができる。

SI08 I can give personal information (address, telephone number, nationality, age, family and hobbies).

個人的な情報（住所、電話番号、国籍、年齢、家族及び趣味）を提供できる。

SI09 I can make simple transactions in shops, post offices or banks.

店・郵便局及び銀行での簡単な手続きができる。

SI10 I can use public transport: buses, trains, and taxis, ask for basic information and buy tickets.

バス・電車及びタクシーといった公共交通機関を使用でき、基本的な情報について尋ねたり切符を購入したりすることができる。

SI11 I can get simple information about travel.

旅行に関する簡単な情報を入手できる。

SI13 I can make simple purchases by stating what I want and asking the price.

欲しい物を伝えたり、価格を尋ねたりして、簡単な買い物をすることができる。

SI14 I can ask for and give directions referring to a map or plan.

地図を見せたり、自分の計画について話したりしながら、道を尋ねることができる。

SI15 I can ask how people are and react to news.

他の人の状況を尋ねたり、ニュースに対して反応したりすることができる。

SI16 I can make and respond to invitations.

招待したり、招待に応じたりすることができる。

SI17 I can make and accept apologies.

謝罪をしたり、謝罪を受け入れたりすることができる。

SI18 I can say what I like and dislike.

好きなこと、嫌いなことが言える。

SI19 I can discuss with other people what to do, where to go and make arrangements to meet.

他の人と何をするか、どこに行くかを話し合い、会う準備をすることができる。

SI20 I can ask people questions about what they do at work and in free time, and answer such questions addressed to me.

他の人が仕事中や余暇に何をしているのか質問することができ、また、そのような質問に答えることができる。

SI21 I can ask for and follow detailed directions.

細かい指示を求め、またそれに従うことができる。

SI22 I can start, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest.

身近なことがらや個人的な興味について、一対一の状態で簡単な会話を始めたり、続けたり、終わらせたりすることができる。

SI23 I can maintain a conversation or discussion but may sometimes be difficult to follow when trying to say exactly what I would like to do.

自分の言いたいことを正確に言おうとすると、ついていくのが困難なことがあるが、会話や議論を維持することができる。

SI24 I can deal with most situations likely to arise when making travel arrangements through an agent or when actually traveling.

旅行会社を通して旅行手続きを行うときや、実際に旅行しているときに起こりうるほとんどの状況に対処できる。

SI25 I can express and respond to feelings such as surprise, happiness, interest and indifference.

驚き、喜び、興味あるいは無関心といった感情を表現し、またこれらに対して反応することができる。

SI26 I can give or seek personal views and opinions in an informal discussion with friends.

友達との形式ばらない議論において、個人的な見解や意見を述べたり尋ねたりすることができる。

SI27 I can agree and disagree politely.

丁寧に賛成したり、反対したりすることができる。

SI28 I can initiate, maintain and end discourse naturally with effective turn-taking.

話し手・聞き手の役割交代を効果的行いながら、会話を自然に開始・維持・終了することができる。

SI29 I can exchange considerable quantities of detailed factual information on matters within my fields of interest.

自分の関心のある分野で、事実についての大量で詳しい情報を交換できる。

SI30 I can convey degrees of emotion and highlight the personal significance of events and experiences.

感情の度合いを伝えることができ、出来事や経験が自分にとってどれだけ重要かを強調することができる。

SI31 I can engage in extended conversation in a clearly participatory fashion on most

general topics.

自分が発言者として参加することが明らかにわかっている状況で、一般的な話題について幅広い会話をすることができる。

SI32 I can account for and sustain my opinions in discussion by providing relevant explanations, arguments and comments.

討論の際に、適切な説明、論拠、コメントを提供することで、自分の意見を述べて正当化することができる。

SI33 I can help a discussion along on familiar ground confirming comprehension, inviting others in, etc.

馴染みのある領域について、他の人の理解を確認したり、参加をうながしたりすることで、討論が進むように手助けできる。

SI34 I can carry out a prepared interview, checking and confirming information, following up interesting replies.

情報を調べたり、確認したり、興味深い返答に対応したりしながら、前もって用意した内容のインタビューをすることができる。

SI35 I can keep up with an animated conversation between native speakers.

ネイティブスピーカー同士のいきいきとした会話についていくことができる。

SI36 I can use the language fluently, accurately and effectively on a wide range of general, professional or academic topics.

一般的、専門的または学問的な幅広い話題において、流暢で正確かつ効果的に言葉を使うことができる。

SI37 I can use language flexibly and effectively for social purposes, including emotional, allusive and joking usage.

人との付き合いの中で、感情表現やほのめかし、冗談を含む言い回しを柔軟かつ効果的に使うことができる。

SI38 I can express my ideas and opinions clearly and precisely, and can present and respond to complex lines of reasoning convincingly.

自分の意見を明瞭かつ簡潔に表現することができ、説得力のある方法で複雑な流れの推論を提案したり反応したりすることができる。

SI39 I can take part effortlessly in all conversations and discussions with native speakers.

ネイティブスピーカーとのあらゆる会話や討論に、努力することなしに参加することができる。

SP01 I can indicate time by such phrases as “next week”, “last Friday”, “in November”, “three o’clock”.

「来週」「先週の金曜日」「11月に」「3時」などといったフレーズを使って、時間を示すことができる。

SP02 I can describe myself, my family and other people.

自分自身や家族及び他の人々について説明できる。

SP03 I can describe where I live.

自分の住んでいるところについて説明できる。

SP04 I can give short, basic descriptions of events.

出来事について、短い基本的な説明ができる。

SP05 I can describe my educational background, my present or most recent job.

自分の学歴や、現在または最近の仕事について説明できる。

SP06 I can describe my hobbies and interests in a simple way.

自分の趣味や関心について簡単な説明ができる。

SP07 I can describe past activities and personal experiences (e.g. the last weekend, my last holiday).

過去の活動や個人的な経験（先週末、この前の休暇など）について説明できる。

SP08 I can narrate a story.

物語を話すことができる。

SP09 I can give detailed accounts of experiences, describing feelings and reactions.

経験したことを、感情や自分の対応をまじえながら詳しく説明することができる。

SP10 I can describe dreams, hopes and ambitions.

夢や希望、野心について述べることができる。

SP11 I can explain and give reasons for my plans, intentions and actions.

自分の計画、意図、行動を説明し、それらの理由を説明することができる。

SP12 I can relate the plot of a book or film and describe my reactions.

本や映画の筋書きを関連付けることができ、自分がどう思ったか述べることができる。

SP13 I can paraphrase short written passages orally in a simple fashion, using the original text wording and ordering.

元の言葉遣いや順序立てを使って、短い段落の文章を簡単に言い換えることができる。

SP14 I can give clear, detailed description on a wide range of subjects related to my fields of interest.

関心のある分野の様々な話題について、わかりやすく詳しい説明をすることができる。

SP15 I can understand and summarize orally short extracts from news items, interviews or documentaries containing opinions, argument and discussion.

短いニュース、インタビュー、意見や議論・討論を含むドキュメンタリーからの抜粋を理解することができ、また口頭で要約することができる。

SP16 I can understand and summarize orally the plot and sequence of events in an extract

from a film or play.

映画や演劇のあらすじや、そこで起こった出来事の流れを理解でき、口頭で要約することができる。

SP17 I can construct a chain of reasoned argument, linking my ideas logically.

自分の考えを論理的に結びつけながら、理路整然とした議論を組み立てることができる。

SP18 I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.

いくつかのことがらに関する有利な点や不利な点をあげながら、話題となっている問題に対する自分の見方を説明できる。

SP19 I can speculate about causes, consequences, hypothetical situations.

会話の中で、原因、結果、仮定されていることなどを推測することができる。

SP20 I can give clear, detailed descriptions of complex subjects.

複雑なテーマについて、明瞭で詳しい説明をすることができる。

SP21 I can orally summarize long, demanding texts.

長くて読みこなすのが大変なテキストを口頭で要約できる。

SP22 I can give an extended description or account of something, integrating themes, developing particular points and concluding appropriately.

何らかのことがらに関して、テーマについてまとめたり、ある点を発展させたり、適切な結論づけを行ったりして、更に進んだ描写や説明をすることができる。

SP23 I can give a clearly developed presentation in my fields of personal or professional interest, departing when necessary from the prepared text and responding to points raised by the audience.

個人的あるいは職業的な関心のある分野に関して、明瞭に組立てられたプレ

ゼンテーションを行うことができる。 また、必要があれば準備した原稿から離れたり、聴衆から指摘された点に臨機応変に対応することができる。

SP24 I can present ideas and viewpoints in a very flexible manner in order to give emphasis, to differentiate and to eliminate ambiguity.

柔軟な態度で強調、区別、あいまいさの排除を行って、アイディアや意見を提示することができる。

SP25 I can summarize orally information from different sources, reconstructing arguments and accounts in a coherent presentation.

さまざまな情報源から得られる情報を口頭で要約し、複数の論点や説明を理路整然としたプレゼンテーションにまとめ直すことができる。

LS01 I can say when I don’t understand.

わからない場合は、わからないと英語で言うことができる。

LS02 I can very simply ask somebody to speak more slowly.

もっとゆっくり言ってくださいと英語で言うことができる。

LS03 I can ask for attention.

英語で相手の注意を引くことができる。

LS04 I can indicate when I am following.

相手の言っていることがわかっているときに、わかっていると伝えることができる。

LS05 I can very simply ask somebody to repeat what they said.

もう一度繰り返して言ってくださいと、英語で言うことができる。

LS06 I can repeat back part of what someone has said to confirm that we understand each other.

お互いの理解を確かめるために相手の言ったことを部分的にくり返すことができる。

LS07 I can ask someone to clarify or elaborate what they have just said.

今言ったことを、明確にまたは詳しく述べてくれるよう頼むことができる。

LS08 When I can't think of the word I want, I can use a simple word meaning something similar and invite “correction”.

言いたい言葉が思い浮かばない場合、同じような意味の簡単な単語を出して、相手に「訂正」してもらうことができる。

LS09 I can use standard phrases like “That’s a difficult question to answer” to gain time and keep the turn while formulating what to say.

話すことを考えている間に時間を稼いで場をもたせるために ”That’s a difficult question to answer” といったような決まり文句を使うことができる。

LS10 I can make a note of "favorite mistakes" and consciously monitor speech for them.

「自分がよくしてしまう間違い」を心に留めておき、会話をしている間意識的に注意できる。

LS11 I can generally correct slips and errors if I become aware of them or if they have led to misunderstandings.

間違いによって誤解が生じたり、自分の間違いに気づいたりした場合は、だいたい、その間違いを直すことができる。

LS12 I can use fluently a variety of appropriate expressions to preface my remarks in order to get the floor, or to gain time and keep the floor while thinking.

発言権を得たり、自分が考えている間に時間をかせいだり、議論を持続させたりするために、場面にふさわしいさまざまな表現の前置きを流暢に使うことができる。

LS13 I can relate own contribution skillfully to those of other speakers.

自分の発言を他の話し手の発言にうまく関連づけることができる。

LS14 I can substitute an equivalent term for a word I can’t recall without distracting the

listener.

聞き手の気をそらさずに、自分が思い出せない単語を別の単語で言い換えることができる。

LS15 I can backtrack and restructure around a difficulty so smoothly the interlocutor is hardly aware of it.

他の話者にほとんど気づかれることなく、前言を撤回したり、難しい部分をスムーズに言いなおしたりすることができる。

LQ01 I can make myself understood using memorised phrases and single expressions.

覚えたフレーズや表現を使って自分の意志を伝えることができる。

LQ02 I can link groups of words with simple connectors like "and", "but", and "because".

and や but、because などの接続詞を用いた簡単な文を使うことができる。

LQ03 I can use some simple structures correctly.

簡単な英文であれば、正しく使うことができる。

LQ04 I have a sufficient vocabulary for coping with simple everyday situations.

簡単で日常的な状況に対処するのに十分なボキャブラリーがある。

LQ05 I can keep a conversation going comprehensibly, but have to pause to plan and correct what I am saying -especially when I talk freely for longer periods.

特にフリートークを長時間している場合、言おうとしていることを考えたり訂正したりするために会話を一時中断する必要があるが、会話を理解しながら続けることができる。

LQ06 I can convey simple information of immediate relevance, getting across which point I feel is most important.

自分がどの点を最も重要だと考えているかを伝え、直接関連している簡単な情報を伝えることができる。

LQ07 I have a sufficient vocabulary to express myself with some circumlocutions on most

topics pertinent to my everyday life such as family, hobbies and interests, work, travel, and current events.

家族、趣味や興味、仕事、旅行や時事などの自分の日常生活に関連のある話題に関して、遠まわしな表現を用いて、自分が言いたいことを言い表す十分なボキャブラリーがある。

LQ08 I can express myself reasonably accurately in familiar, predictable situations.

馴染みのある予測可能な状況では、自分の考えをほぼ正確に伝えることができる。

LQ09 I can produce stretches of language with a fairly even tempo; although I can be hesitant as I search for expressions, there are few noticeably long pauses.

表現を探すのに少し口ごもるけれども、めだった長いポーズがなく、一定の速度で言葉にすることができる。

LQ10 I can pass on detailed information reliably.

詳しい情報を確実に伝えることができる。

LQ11 I have sufficient vocabulary to express myself on matters connected to my field and on most general topics.

自分の専門に関することや一般的なトピックについて、自分の言いたいことを言うのに十分なボキャブラリーがある。

LQ12 I can communicate with reasonable accuracy and can correct mistakes if they have led to misunderstandings.

ある程度正確にコミュニケーションをとることができ、誤解が生じた場合は間違いを直すことができる。

LQ13 I can express myself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.

概念的に難しい問題に関してのみ、自然でスムーズに述べることができない

が、自分の言いたいことを流暢かつ自然に、特に意識せずに表現できる。

LQ14 I can produce clear, smoothly-flowing, well-structured speech, showing control over ways of developing what I want to say in order to link both my ideas and my expression of them into coherent text.

言うべきことをうまくコントロールしながら、自分の考えとそれを表す表現を結びつけて一貫した文章を作り、はっきりと滑らかに理路整然と話すことができる。

LQ15 I have a good command of a broad vocabulary allowing gaps to be readily overcome with circumlocutions; I rarely have to search obviously for expressions or compromise on saying exactly what I want to.

あからさまに表現を探したり、自分が言いたいことを正確に言うのをあきらめたりすることはめったになく、知らない単語でも知っている単語で補えるだけの幅広い語彙を自由に使いこなすことができる。

LQ16 I can consistently maintain a high degree of grammatical accuracy; errors are rare and difficult to spot.

誤りはほとんどなく、誤りを見抜くことも難しいくらい、文法的な正確性を一貫して維持できる。

LQ17 I can express myself naturally and effortlessly; I only need to pause occasionally in order to select precisely the right words.

正しい言葉を正確に選択するために時間を要することも時々あるが、自然に苦労することなく自分の意見を述べることができる。

LQ18 I can convey finer shades of meaning precisely by using, with reasonable accuracy, a wide range of expressions to qualify statements and pinpoint the extent to which something is the case.

表現を和らげたりものごとの程度を正確に示すために様々な表現をおおよそ

正しく使用し、細かい意味のニュアンスを的確に伝えることができる。

LQ19　I have a good command of idiomatic expressions and colloquialisms with an awareness of implied meaning and meaning by association.

間接的な意味や連想を意識して、英語らしい表現やこなれた表現を使うことができる。

LQ20　I can consistently maintain grammatical control of complex language even when my attention is otherwise engaged.

気が散っているときでも、常に複雑な文法を使いこなすことができる。

Appendix B: Passing rates and point biserial correlation coefficients, and the difficulty and the discrimination power calculated by IRT

Spoken Interaction (students)

	# of response	Passing rate	p.b.s	discrimination	Difficulty
1	449	0.911	0.611	1.200	-1.631
2	449	0.864	0.729	1.636	-1.166
3	449	0.924	0.599	1.223	-1.718
4	449	0.833	0.472	0.828	-1.472
5	449	0.706	0.337	0.556	-1.110
6	449	0.690	0.651	1.281	-0.653
7	449	0.693	0.678	1.386	-0.641
8	449	0.673	0.561	0.958	-0.677
9	1040	0.212	0.479	0.786	1.153
10	1040	0.605	0.469	1.056	-0.399
11	1040	0.613	0.509	1.051	-0.426
12	1040	0.911	0.505	1.139	-1.631
13	1040	0.871	0.550	1.350	-1.250
14	1040	0.685	0.516	1.019	-0.662
15	1040	0.495	0.571	0.992	-0.071
16	1040	0.571	0.619	1.095	-0.295
17	1040	0.726	0.529	1.004	-0.812
18	1040	0.923	0.500	1.121	-1.749
19	1040	0.744	0.668	1.383	-0.747

20	1040	0.712	0.614	1.285	-0.674
21	1641	0.760	0.514	1.557	-0.590
22	1641	0.568	0.550	1.481	-0.095
23	1641	0.208	0.518	0.959	1.152
24	1641	0.399	0.542	1.152	0.359
25	1641	0.779	0.480	1.175	-0.769
26	1641	0.733	0.546	1.546	-0.513
27	1641	0.627	0.536	1.170	-0.279
28	2170	0.717	0.548	1.497	-0.156
29	2170	0.547	0.509	1.191	0.270
30	2170	0.594	0.567	1.097	0.131
31	2170	0.567	0.580	1.591	0.232
32	2170	0.508	0.592	1.560	0.371
33	2170	0.624	0.494	1.275	0.068
34	2170	0.606	0.508	1.031	0.084
35	1579	0.381	0.627	1.549	0.835
36	1579	0.284	0.723	1.811	1.059
37	1579	0.479	0.537	1.099	0.633
38	1579	0.405	0.704	1.970	0.736
39	978	0.366	0.696	1.574	1.031

Spoken Production (student)

	# of response	Passing rate	p.b.s	discrimination	Difficulty
1	449	0.822	0.444	0.687	-1.596

2	1040	0.875	0.552	1.048	-1.421
3	1040	0.868	0.587	1.150	-1.298
4	1040	0.816	0.595	1.271	-0.964
5	1040	0.769	0.596	1.408	-0.735
6	1040	0.885	0.603	1.099	-1.444
7	1040	0.829	0.586	1.199	-1.056
8	1641	0.539	0.371	1.066	0.112
9	1641	0.640	0.531	1.347	-0.164
10	1641	0.725	0.545	0.988	-0.550
11	1641	0.683	0.594	1.252	-0.311
12	1641	0.445	0.482	1.154	0.396
13	1641	0.416	0.505	1.131	0.492
14	2170	0.585	0.528	1.104	0.364
15	2170	0.397	0.603	1.230	0.963
16	2170	0.457	0.553	1.222	0.772
17	2170	0.375	0.641	1.358	1.012
18	2170	0.555	0.554	1.317	0.477
19	2170	0.637	0.470	1.069	0.186
20	1579	0.311	0.769	1.635	1.356
21	1579	0.367	0.676	1.287	1.238
22	1579	0.405	0.759	1.739	1.078
23	1579	0.392	0.695	1.329	1.153
24	978	0.516	0.735	1.527	0.967
25	978	0.417	0.824	2.323	1.182

Language Strategies (student)

	# of response	Passing rate	p.b.s	discrimination	Difficulty
1	449	0.457	0.608	2.099	-0.049
2	449	0.445	0.600	2.605	-0.044
3	1040	0.407	0.381	0.315	1.472
4	1040	0.677	0.635	0.970	-0.124
5	1040	0.698	0.660	2.550	-0.083
6	1641	0.683	0.619	0.688	0.108
7	1641	0.701	0.483	0.578	-0.092
8	1641	0.620	0.552	0.523	0.368
9	2170	0.555	0.466	0.530	1.721
10	2170	0.547	0.446	0.534	1.769
11	2170	0.679	0.467	0.617	1.032
12	1579	0.413	0.590	0.783	2.747
13	1579	0.646	0.562	0.753	1.618
14	1579	0.503	0.619	0.770	2.308
15	978	0.383	0.623	0.807	3.193

Language Quality (student)

	# of response	Passing rate	p.b.s	discrimination	Difficulty
1	1040	0.748	0.578	0.984	-0.955
2	1040	0.862	0.502	0.878	-1.646
3	1040	0.788	0.518	0.889	-1.200
4	1040	0.424	0.579	1.015	0.183

5	1641	0.595	0.567	1.100	-0.287
6	1641	0.552	0.576	1.339	-0.138
7	1641	0.382	0.617	1.270	0.339
8	1641	0.604	0.581	1.405	-0.275
9	2170	0.403	0.670	1.307	0.447
10	2170	0.369	0.700	1.641	0.505
11	2170	0.361	0.658	1.331	0.567
12	2170	0.659	0.590	1.327	-0.277
13	1579	0.477	0.754	1.948	0.322
14	1579	0.407	0.816	2.698	0.449
15	1579	0.434	0.755	1.872	0.419
16	1579	0.247	0.725	1.359	1.030
17	978	0.553	0.765	1.742	0.241
18	978	0.424	0.768	1.627	0.559
19	978	0.403	0.791	1.707	0.608
20	978	0.310	0.773	1.496	0.903

Spoken Interaction (teacher)

	# of response	Passing rate	p.b.s	discrimination	Difficulty
1	166	0.892	0.950	2.237	-1.167
2	166	0.892	0.930	2.103	-1.190
3	166	0.789	0.696	1.292	-0.968
4	166	0.898	1.011	2.402	-1.171
5	166	0.825	0.786	1.533	-1.036

6	166	0.831	0.878	2.009	-0.966
7	166	0.904	1.046	3.199	-1.121
8	166	0.922	0.945	2.051	-1.369
9	391	0.772	0.800	1.712	-0.670
10	391	0.752	0.738	1.421	-0.654
11	391	0.816	0.915	2.615	-0.715
12	391	0.928	0.846	1.422	-1.536
13	391	0.926	0.907	1.685	-1.390
14	391	0.680	0.726	1.469	-0.424
15	391	0.821	0.852	1.943	-0.797
16	391	0.818	0.978	2.415	-0.738
17	391	0.744	0.850	1.989	-0.556
18	391	0.941	0.839	1.570	-1.587
19	391	0.818	0.865	2.073	-0.771
20	391	0.821	0.963	3.25	-0.699
21	620	0.863	0.833	2.341	-0.770
22	620	0.748	0.749	2.049	-0.438
23	620	0.553	0.628	1.744	0.036
24	620	0.658	0.673	1.780	-0.222
25	620	0.790	0.785	1.873	-0.579
26	620	0.768	0.772	2.726	-0.446
27	620	0.806	0.654	1.725	-0.656
28	816	0.797	0.803	2.331	-0.303
29	816	0.678	0.736	2.241	0.008

30	816	0.721	0.849	2.894	-0.073
31	816	0.743	0.832	2.767	-0.130
32	816	0.669	0.794	2.777	0.043
33	816	0.688	0.816	2.253	-0.015
34	816	0.510	0.628	0.981	0.429
35	591	0.425	0.715	1.799	0.702
36	591	0.475	0.770	2.147	0.567
37	591	0.489	0.771	2.132	0.536
38	591	0.518	0.791	2.230	0.469
39	362	0.475	0.803	2.092	0.730

Spoken Production (teacher)

	# of response	Passing rate	p.b.s	discrimination	Difficulty
1	528	0.271	0.415	2.392	-1.082
2	391	0.895	0.864	2.388	-1.008
3	391	0.921	0.825	2.039	-1.232
4	391	0.900	0.799	1.831	-1.138
5	391	0.808	0.723	1.823	-0.709
6	391	0.954	0.886	2.147	-1.538
7	391	0.900	0.839	2.062	-1.088
8	620	0.581	0.558	1.760	0.061
9	620	0.608	0.695	2.082	0.007
10	620	0.737	0.732	1.746	-0.362
11	620	0.771	0.653	1.86	-0.454

12	620	0.474	0.560	2.055	0.313
13	620	0.394	0.658	1.616	0.536
14	816	0.619	0.616	1.722	0.302
15	816	0.376	0.692	1.982	0.899
16	816	0.477	0.630	1.557	0.665
17	816	0.603	0.670	1.797	0.348
18	816	0.588	0.583	1.848	0.388
19	816	0.548	0.564	1.852	0.488
20	591	0.437	0.739	1.972	0.857
21	591	0.223	0.766	1.789	1.468
22	591	0.396	0.703	1.897	0.952
23	591	0.313	0.718	1.361	1.25
24	362	0.45	0.774	2.232	0.988
25	362	0.37	0.799	2.301	1.145

Language Strategies (teacher)

	# of response	Passing rate	p.b.s	discrimination	Difficulty
1	166	0.831	0.983	3.028	-0.985
2	166	0.825	0.995	3.154	-0.958
3	391	0.721	0.754	1.649	-0.614
4	391	0.760	0.882	2.472	-0.664
5	391	0.867	0.851	2.062	-1.076
6	620	0.715	0.752	2.206	-0.464
7	620	0.689	0.718	1.906	-0.412

8	620	0.613	0.714	1.850	-0.213
9	816	0.596	0.663	1.473	0.055
10	816	0.475	0.683	1.448	0.397
11	816	0.694	0.697	1.572	-0.224
12	591	0.431	0.781	1.960	0.557
13	591	0.516	0.712	1.508	0.353
14	591	0.509	0.773	2.020	0.366
15	362	0.373	0.736	1.967	0.821

Language Quality (teacher)

	# of response	Passing rate	p.b.s	discrimination	Difficulty
1	391	0.818	0.820	1.493	-0.905
2	391	0.839	0.788	1.703	-0.941
3	391	0.898	0.928	2.216	-1.140
4	391	0.806	0.845	1.925	-0.769
5	620	0.705	0.777	1.756	-0.394
6	620	0.732	0.743	1.882	-0.462
7	620	0.629	0.777	2.122	-0.170
8	620	0.784	0.769	2.023	-0.608
9	816	0.600	0.747	2.094	0.077
10	816	0.630	0.692	2.253	0.019
11	816	0.713	0.649	2.260	-0.175
12	816	0.730	0.755	2.633	-0.194
13	591	0.420	0.851	3.033	0.461

14	591	0.415	0.833	3.280	0.465
15	591	0.411	0.850	4.195	0.458
16	591	0.431	0.810	3.052	0.439
17	362	0.489	0.936	4.269	0.430
18	362	0.384	0.852	3.203	0.601
19	362	0.354	0.784	2.338	0.692
20	362	0.376	0.879	3.681	0.605

Appendix C: Details of measurement report of the evaluation of spontaneous speech

Teacher measurement report (first analysis)

	Obsrvd Average	Fair-M (logits)	Severity	Error	Infit	Outfit	Estim. Discrm.
Rater 1	3.0	2.99	0.53	0.03	1.16	1.15	0.84
Rater 2	4.1	4.22	-0.78	0.03	0.88	0.88	1.12
Rater 3	3.3	3.03	0.19	0.03	0.53	0.57	1.51
Rater 4	3.6	3.66	-0.19	0.03	0.65	0.65	1.38
Rater 5	4.0	4.06	-0.61	0.03	1.07	1.07	0.89
Rater 6	3.7	3.76	-0.29	0.03	1.37	1.39	0.59
Rater 7	3.7	3.74	-0.27	0.03	1.01	1.03	1.01
Rater 8	3.4	3.44	0.05	0.03	0.47	0.49	1.54
Rater 9	3.9	3.89	-0.43	0.03	1.26	1.27	0.70
Rater 10	4.7	4.84	-1.55	0.03	1.65	1.54	0.38
Mean	3.8	3.79	-0.34	0.03	1.00	1.00	
S.D.	0.4	0.52	0.57	0.00	0.38	0.35	

Rater measurement report (Second analysis)

	Obsrvd Average	Fair-M (logits)	Severity	Error	Infit	Outfit	Estim. Discrm.
Rater 1	3	2.99	0.56	0.03	1.16	1.15	0.84
Rater 2	4.2	4.28	-0.92	0.03	0.87	0.88	1.12
Rater 3	3.3	3.28	0.22	0.03	0.48	0.49	1.57
Rater 4	3.7	3.66	-0.22	0.03	0.64	0.64	1.39
Rater 5	4.1	4.12	-0.74	0.03	1.13	1.13	0.81
Rater 6	3.8	3.81	-0.38	0.03	1.38	1.41	0.57
Rater 7	3.7	3.76	-0.33	0.03	1.03	1.04	1.00
Rater 8	3.5	3.45	0.02	0.03	0.49	0.5	1.53
Rater 9	4.0	4.00	-0.6	0.03	1.25	1.26	0.72
Rater 10	4.9	4.97	-1.86	0.03	1.62	1.54	0.4
Mean	3.8	3.83	-0.43	0.03	1.01	1.00	
S.D.	0.5	0.56	0.67	0	0.38	0.37	

Rater measurement report (Third analysis)

	Obsrvd Average	Fair-M (logits)	Severity	Error	Infit	Outfit	Estim. Discrm.
Rater 1	3.0	2.95	0.65	0.03	1.14	1.14	0.85
Rater 2	4.2	4.30	-0.97	0.03	0.87	0.88	1.12
Rater 3	3.3	3.28	0.24	0.03	0.49	0.5	1.55
Rater 4	3.7	3.67	-0.22	0.03	0.65	0.65	1.38
Rater 5	4.1	4.11	-0.74	0.03	1.07	1.08	0.89
Rater 6	3.8	3.81	-0.38	0.03	1.43	1.46	0.51
Rater 7	3.7	3.77	-0.34	0.03	1.02	1.03	1.01
Rater 8	3.5	3.46	0.02	0.03	0.51	0.52	1.51
Rater 9	4.0	4.01	-0.62	0.03	1.22	1.23	0.75
Rater 10	4.9	5.00	-1.99	0.03	1.64	1.55	0.38
Mean	3.8	3.84	-0.44	0.03	1.00	1.00	
S.D.	0.5	0.57	0.73	0.00	0.38	0.37	

Item 1: Loudness

Item 2: Sound pitch

Item 3: Quality of vowels

Item 4: Quality of consonants

Item 5: Epenthesis

Item 6: Elision

Item 7: Word stress

Item 8: Sentence stress

Item 9: Speech rate

Item 10: Prosody

Item 11: Fluency

Item 12: Place of fillers

Item 13: Frequency of fillers

Item 14: Place of silent pause

Item 15: Frequency of silent pause

Item 16: Length of silent pause

Item 17: Relevant paralinguistic cues

Item: 18: Confidence

Item: 19: Try to sound cheerful

Item: 20: Try to sound friendly

Item: 21: Grammatical Accuracy

Item: 22: Coherency

Item 23: Absence of tension

Item 24: Foreign accentednes

Item measurement report (First analysis)

	Obsrvd Average	Fair-M (logits)	Difficulty	Error	Infit	Outfit	Estim. Discrm.
Item 1	4.4	4.50	-0.77	0.04	1.05	1.10	0.92
Item 2	4.4	4.53	-0.81	0.04	0.92	0.96	1.05
Item 3	3.7	3.71	0.10	0.04	0.79	0.81	1.20
Item 4	3.7	3.69	0.11	0.04	0.81	0.85	1.15
Item 5	3.7	3.71	0.09	0.04	1.00	1.02	0.98

Item 6	3.7	3.77	0.03	0.04	0.92	0.94	1.06
Item 7	3.9	3.98	-0.19	0.04	0.67	0.69	1.36
Item 8	3.8	3.85	-0.05	0.04	0.70	0.70	1.35
Item 9	3.7	3.69	0.11	0.04	0.85	0.86	1.16
Item 10	3.9	3.98	-0.18	0.04	0.79	0.79	1.25
Item 11	3.7	3.7	0.11	0.04	0.97	0.95	1.06
Item 12	3.6	3.66	0.14	0.04	0.91	0.9	1.11
Item 13	3.5	3.48	0.34	0.04	0.98	0.98	1.01
Item 14	3.7	3.74	0.07	0.04	0.94	0.94	1.09
Item 15	3.6	3.62	0.19	0.04	0.96	0.96	1.07
Item 16	3.7	3.7	0.11	0.04	0.99	0.98	1.04
Item 17	3.8	3.87	-0.07	0.04	1.35	1.35	0.61
Item 18	3.8	3.80	0.00	0.04	1.21	1.20	0.79
Item 19	3.9	3.97	-0.18	0.04	0.83	0.86	1.15
Item 20	4.0	4.05	-0.27	0.04	0.74	0.79	1.24
Item 21	3.8	3.88	-0.08	0.04	0.85	0.85	1.18
Item 22	3.8	3.79	0.01	0.04	1.02	1.03	1.00
Item 23	3.6	3.58	0.23	0.04	1.68	1.69	0.23
Item 24	3	2.92	0.95	0.04	1.90	1.89	-0.06
Mean	3.8	3.8	0.00	0.04	0.99	1.00	
S.D.	0.2	0.31	0.34	0.00	0.29	0.28	

Item measurement report (Second analysis)

	Obsrvd Average	Fair-M (logits)	Difficulty	Error	Infit	Outfit	Estim. Discrm.
Item 1	4.4	4.5	-0.77	0.04	1.11	1.16	0.85
Item 2	4.4	4.53	-0.81	0.04	0.97	1.01	0.99
Item 3	3.7	3.70	0.16	0.04	0.86	0.88	1.12
Item 4	3.7	3.69	0.18	0.04	0.89	0.94	1.05
Item 5	3.7	3.71	0.15	0.04	1.07	1.09	0.9
Item 6	3.7	3.76	0.09	0.04	0.99	1.02	0.98
Item 7	3.9	3.98	-0.15	0.04	0.72	0.73	1.30
Item 8	3.8	3.84	0.00	0.04	0.75	0.75	1.29
Item 9	3.7	3.69	0.18	0.04	0.93	0.94	1.07
Item 10	3.9	3.97	-0.14	0.04	0.85	0.85	1.18
Item 11	3.7	3.69	0.17	0.04	1.07	1.05	0.94
Item 12	3.6	3.66	0.21	0.04	0.94	0.94	1.07
Item 13	3.5	3.47	0.42	0.04	1.05	1.05	0.93
Item 14	3.7	3.73	0.13	0.04	0.99	0.98	1.03
Item 15	3.6	3.61	0.26	0.04	1.03	1.03	0.99
Item 16	3.7	3.69	0.17	0.04	1.06	1.05	0.96
Item 17	3.8	3.86	-0.02	0.04	1.46	1.47	0.49
Item 18	3.8	3.79	0.06	0.04	1.34	1.32	0.64
Item 19	3.9	3.96	-0.13	0.04	0.90	0.94	1.07
Item 20	4.0	4.05	-0.23	0.04	0.80	0.84	1.17
Item 21	3.8	3.88	-0.03	0.04	0.91	0.91	1.10

Item 22	3.8	3.79	0.07	0.04	1.11	1.13	0.89
Mean	3.8	3.84	0.00	0.04	0.99	1.00	
S.D.	0.2	0.26	0.30	0.00	0.17	0.17	

Item measurement report (Third analysis)

	Obsrvd Average	Fair-M (logits)	Difficulty	Error	Infit	Outfit	Estim. Discrm.
Item 1	4.4	4.50	-0.80	0.04	1.15	1.20	0.82
Item 2	4.4	4.53	-0.84	0.04	1.00	1.04	0.96
Item 3	3.7	3.71	0.17	0.04	0.87	0.89	1.11
Item 4	3.7	3.69	0.19	0.04	0.91	0.97	1.03
Item 5	3.7	3.71	0.16	0.04	1.10	1.12	0.88
Item 6	3.7	3.77	0.10	0.04	1.02	1.04	0.95
Item 7	3.9	3.98	-0.15	0.04	0.74	0.75	1.28
Item 8	3.8	3.85	0.01	0.04	0.76	0.76	1.27
Item 9	3.7	3.69	0.19	0.04	0.96	0.97	1.04
Item 10	3.9	3.97	-0.14	0.04	0.89	0.89	1.13
Item 11	3.7	3.70	0.18	0.04	1.14	1.13	0.86
Item 12	3.6	3.66	0.22	0.04	0.98	0.97	1.03
Item 13	3.5	3.48	0.44	0.04	1.09	1.10	0.88
Item 14	3.7	3.73	0.14	0.04	1.04	1.03	0.97
Item 15	3.6	3.62	0.28	0.04	1.10	1.09	0.91
Item 16	3.7	3.70	0.18	0.04	1.13	1.12	0.88
Item 19	3.9	3.97	-0.13	0.04	0.95	0.99	1.02

Item 20	4.0	4.05	-0.24	0.04	0.83	0.88	1.14
Mean	3.8	3.88	-0.03	0.04	0.95	0.95	1.06
S.D.	3.8	3.79	0.07	0.04	1.18	1.20	0.81

Appendix D: Details of measurement report of raters and items in the evaluation of read-aloud speech

Rater measurement report

	Obsrvd Average	Fair-M (logits)	Severity	Error	Infit	Outfit	Estim. Discrm.
Rater 1	4.1	4.19	-0.82	0.03	1.30	1.28	0.66
Rater 2	3.8	3.84	-0.40	0.03	0.84	0.87	1.13
Rater 3	4.2	4.23	-0.87	0.03	1.00	0.99	1.04
Rater 4	3.6	3.64	-0.17	0.03	0.96	0.96	1.06
Rater 5	3.2	3.23	0.31	0.03	0.68	0.69	1.32
Rater 6	4.2	4.25	-0.9	0.03	1.21	1.2	0.79
Mean	3.9	3.9	-0.48	0.03	1.00	1.00	
S.D.	0.3	0.41	0.48	0.00	0.21	0.22	

Item measurement report

	Obsrvd Average	Fair-M (logits)	Difficulty	Error	Infit	Outfit	Estim. Discrm.
Item 1	4.3	4.42	-0.65	0.05	1.00	1.05	0.94
Item 2	4.4	4.5	-0.76	0.05	0.95	0.97	1.01
Item 3	3.6	3.58	0.38	0.04	0.97	0.97	1.03
Item 4	3.7	3.73	0.21	0.04	0.94	0.94	1.09
Item 5	3.8	3.86	0.05	0.04	1.14	1.13	0.88
Item 6	3.7	3.71	0.23	0.04	1.17	1.16	0.80
Item 7	4	4.03	-0.15	0.04	1.15	1.13	0.84

Item 8	3.7	3.67	0.27	0.04	0.88	0.88	1.16
Item 9	3.5	3.49	0.47	0.04	1.01	1.01	1.00
Item 10	3.5	3.46	0.51	0.04	0.91	0.91	1.1
Item 11	4.0	4.02	-0.14	0.04	0.95	0.95	1.04
Item 12	4.0	4.02	-0.14	0.04	0.95	0.95	1.04
Item 13	4.0	4.02	-0.14	0.04	0.95	0.95	1.04
Item 14	4	4.02	-0.14	0.04	0.95	0.95	1.04
Mean	3.9	3.9	0.00	0.04	0.99	1.00	
S.D.	0.2	0.3	0.37	0.00	0.09	0.08	

Appendix E: The pronunciation dictionary for HTK

```
a               ah sp
a               ax sp
a               ey sp
aan             ax n sp
agree           ax g r iy sp
agreed          ax g r iy d sp
agrees          ax g r iy z sp
along           ax l oh ng sp
and             ae n d sp
and             ax n sp
and             ax n d sp
around          ax r aw n d sp
as              ae s sp
as              ae z sp
as              ax z sp
at              ae t sp
at              ax t sp
attempt         ax t eh m p t sp
be              b iy sp
beau            b ow sp
below           b ih l ow sp
blew            b l uw sp
blews           b l uw z sp
but             b ah t sp
```

```
but           b ax t sp
came          k ey m sp
can't         k ae n t sp
cloak         k l ow k sp
close         k l ow s sp
close         k l ow z sp
closely       k l ow s l iy sp
closer        k l ow s ax sp
closer        k l ow s ax r sp
closer        k l ow z ax sp
cloth         k l oh th sp
coat          k ow t sp
coke          k ow k sp
come          k ah m sp
con           k oh n sp
confess       k ax n f eh s sp
confessed     k ax n f eh s t sp
cons          k oh n s sp
cons          k oh n z sp
consider      k ax n s ih d ax sp
consider      k ax n s ih d ax r sp
considered    k ax n s ih d ax d sp
could         k uh d sp
dial          d ay ax l sp
did           d ih d sp
```

disputing d ih s p y uw t ih ng sp

diputing d ih p y uw t ih ng sp

earth er th sp

ehh ax sp

eight ey t sp

fastest f aa s t ih s t sp

first f er s t sp

five f ay v sp

flood f l ah d sp

fold f ow l d sp

four f ao sp

four f ao r sp

gave g ey v sp

give g ih v sp

hard hh aa d sp

has hh ae z sp

have ae v sp

have hh ae v sp

he hh iy sp

him hh ih m sp

his hh ih z sp

ill ay l sp

ill ih l sp

im ay m sp

immediately ih m iy d ia t l iy sp

in	ih n sp
is	ay z sp
is	ih z sp
it	ih t sp
just	jh ah s t sp
key	k iy sp
lady	l ey d iy sp
last	l aa s t sp
last	l ae s t sp
least	l iy s t sp
making	m ey k ih ng sp
map	m ae p sp
me	m iy sp
mm	m sp
more	m ao sp
more	m ao r sp
morning	m ao n ih ng sp
nine	n ay n sp
north	n ao th sp
ob	oh b sp
obl	oh b l sp
obli	oh b l iy sp
obligate	oh b l ih g ey t sp
oblige	ax b l ay jh sp
oblige	ax b l iy zh sp

```
obliged        ax b l ay jh d sp
oblique        ax b l iy k sp
of             ax v sp
of             oh v sp
off            ao f sp
off            oh f sp
oh             ow sp
okey           ow k ey sp
one            w ah n sp
other          ah dh ax sp
other          ah dh ax r sp
others         ah dh ax r z sp
others         ah dh ax z sp
out            aw t sp
pen            p eh n sp
rock           r oh k sp
seven          s eh v n sp
shh            sh sp
shine          sh ay n sp
shone          sh oh n sp
shoot          sh uw t sp
shore          sh ao sp
shore          sh ao r sp
should         sh uh d sp
show           sh ow sp
```

```
silence        sil
six            s ih k s sp
so             s ow sp
sorry          s oh r iy sp
stone          s t ow n sp
strong         s t r oh ng sp
stronger       s t r oh ng g ax sp
stronger       s t r oh ng g ax r sp
strongers      s t r oh ng g ax r z sp
suc            s ax k sp
succeed        s ax k s iy d sp
succeeded      s ax k s iy d ih d sp
succeeding     s ax k s iy d ih ng sp
sun            s ah n sp
take           t ey k sp
than           dh ae n sp
that           dh ae t sp
that           dh ax t sp
the            dh ax sp
the            dh iy sp
them           dh eh m sp
then           dh eh n sp
they           dh ey sp
this           dh ih s sp
three          th r iy sp
```

to	t ax sp
to	t uw sp
took	t uh k sp
trav	t r ae v sp
travel	t r ae v l sp
traveler	t r ae v l ax sp
traveler	t r ae v l ax r sp
travelers	t r ae v l ax r z sp
travelers	t r ae v l ax z sp
traven	t r ae v n sp
trong	t r oh ng g sp
two	t uw sp
up	ah p sp
us	ah s sp
us	ah z sp
us	y uw z sp
warm	w ao m sp
warming	w ao m ih ng sp
warmly	w ao m l iy sp
was	w ax z sp
was	w oh z sp
were	w er sp
were	w er r sp
were	w ia sp
were	w ia r sp

were	w iy v sp
what	w oh t sp
when	w eh n sp
which	w ih ch sp
who	hh uw sp
whose	hh uw z sp
wind	w ay n d sp
wind	w ih n d sp
window	w ih n d ow sp
wrap	r ae p sp
wrapped	r ae p t sp
wrapping	r ae p ih ng sp
you	y ax sp
you	y uw sp
zero	z ia r ow sp

Appendix F: Phonetic symbol table for pronunciation dictionary

Example	Phonemic symbol	Symbol in the dictionary
c<u>o</u>nfess	ə	ax, ix
talk	ɔ:	ao
pit	ɪ	ih
peat	i:	iy
but	ʌ	ah
pat	æ	ae
err	ɚ	er
food	u:	uw
lock	ɑ	aa
pen	e, ɛ	eh
put	ʊ	uh
pay	eɪ	ey
lie	aɪ	ay
cloak	oʊ	ow
cow	aʊ	aw

Appendix G: Informed consent

Informed Consent Form

I state that I am over 18 years old and wish to participate in speech recording. I understand that the speech recordings collected here will be used for academic research and analysis only. I agree to the distribution of my recorded speech data for the purpose of research and analysis and understand that such distributions will not involve the identity of the original speakers.

Name of Participant: (English) _____

Native speaker of: _____

Signature of Participant: _____

Date: _____

Appendix H: Perl scripts for controlling the evaluation system

eva.cgi

```
#!/usr/bin/perl -w

use strict;
use CGI;
use List::Util qw/min/;

my ($q, $filename, $eva1, $eva2, $eva3, $eva4, $eva5, $ratio1,
$ratio2, $ratio3, $ratio4, $ratio5, $speed1, $speed2, $speed3,
$speed4, $speed5, $data, $ratio_ave, $speed_ave, $dist, $cate1,
$cate2, $cate3, $min);

$q = new CGI;


$filename = $q->param('filename');

if ( -e "./eva/$filename.5.eva"){

$eva1 = "./eva/$filename.1.eva";
$eva2 = "./eva/$filename.2.eva";
$eva3 = "./eva/$filename.3.eva";
$eva4 = "./eva/$filename.4.eva";
```

```
$eva5 = "./eva/$filename.5.eva";

open (EVA1, "$eva1") or die "Cannot open $eva1 file!¥n";
$data = <EVA1>;
($ratio1, $speed1) = split (/,/, $data);
close (EVA1);

open (EVA2, "$eva2") or die "Cannot open $eva2 file!¥n";
$data = <EVA2>;
($ratio2, $speed2) = split (/,/, $data);
close (EVA2);

open (EVA3, "$eva3") or die "Cannot open $eva3 file!¥n";
$data = <EVA3>;
($ratio3, $speed3) = split (/,/, $data);
close (EVA3);

open (EVA4, "$eva4") or die "Cannot open $eva4 file!¥n";
$data = <EVA4>;
($ratio4, $speed4) = split (/,/, $data);
close (EVA4);

open (EVA5, "$eva5") or die "Cannot open $eva5 file!¥n";
$data = <EVA5>;
($ratio5, $speed5) = split (/,/, $data);
```

```
close (EVA5);

$ratio_ave = (1 - (($ratio1 + $ratio2 + $ratio3 + $ratio4 + $ratio5)
/ 5)) * 10;
$speed_ave = ($speed1 + $speed2 + $speed3 + $speed4 + $speed5) /
5;

$cate1 = sqrt (( $ratio_ave - 4.4 )**2 + ( $speed_ave - 35 )**2);
$cate2 = sqrt (( $ratio_ave - 3.8 )**2 + ( $speed_ave - 31.4 )**2);
$cate3 = sqrt (( $ratio_ave - 3.9 )**2 + ( $speed_ave - 26.7 )**2);


open (BACK,">./back/$filename.back") or die "Cannot open back
file!¥n";
  print BACK "$cate1, $cate2, $cate3";
close (BACK);

my @array = ($cate1, $cate2, $cate3);

$min = List::Util::min( @array );

if ($min == $cate1){
   print"Content-type: text/html; charset=Shift_JIS¥n¥n";
print <<END_OF_HTML;
```

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd">
<html lang=ja>
<head>
<title>音読自動評価 | Evaluation</title>
<STYLE type="text/css">
<!--

#cont {line-height:1em;
padding:0.5em;font-size:13px;font-family:Trebuchet
MS;line-height:17px;padding:0.5em 5em; color:#555555}

A:link { color:#00FFFF; }
      A:visited { color:#6495ED; }
      A:hover { color:#F5FFFA; }
      a { text-decoration: none }

--></style>
<SCRIPT TYPE="text/javascript">
<!--
var sound2Embed = null;
function sound2Play() {
  if ( !sound2Embed ) {
    sound2Embed = document.createElement("embed");
```

```
    sound2Embed.setAttribute("src", "sample.wav");
    sound2Embed.setAttribute("hidden", true);
    sound2Embed.setAttribute("autostart", true);
  } else sound2Stop();
  sound2Embed.removed = false;
  document.body.appendChild(sound2Embed);
}

// -->
</script>

</head>

<BODY bgcolor=#FFF9F9>
<div id="cont">
<h3>evaluation</h3>

<p>このテストでは A, B, C の 3 段階で英語の音読を評価しています。以下があなたの判定結果です。</p>
<p><b>Category A</b>: 微妙な意味もイントネーションなどで¥表¥現でき、明瞭で自然な発音である。</p>

<h3>単語の発音について</h3>
<p>コミュニケーションを阻害するものではありませんが、いくつかの発音が正確ではない可¥能¥性があります。例えば、テキストに出てきた last、sun、at の母音を
```

日本語では同じ「あ」と認識しますが、英語では、これらの母音はすべて異なります。ひとつの母音が異なるだけで意味が異なる場合があります。例えば、but と bat は日本語風に発音すると「バット」になってしまいますが、実際には異なる発音です。また、日本語にない子音についても同様のことが言える可¥能¥性があります。個々の発音に注意を払えば、より伝わりやすい英語になるでしょう。

<h3>文の読み方について</h3>
ほぼ完璧だと思われますが、いくつかの文の強勢が正確ではない可¥能¥性があります。単語にアクセント（強勢）があるように、文にも最も強く発音される単語があります。これは前後関係や話者の意図によってことなりますが、例えば、テキストに出てきた"and at last the North Wind gave up the attempt"という文の場合、一般的に最後の単語 attempt が最も強く発音されます。
ほぼ¥完¥璧だと思われますが、いくつかの文でイントネーション（抑揚）が不適切な可¥能¥性があります。イントネーションの付け方は一通りに決まるものではありませんが、例えば、文の途中でポーズを置く際に文の終りと同じように声の高低を変化させることは適切ではありません。次の例を聞いて下さい。

例

ここではテキストに出てきた文を次のように 3 つに区切って読んでいます。

They agreed ＜ポーズ＞ that one who succeeded in making the traveler take his cloak off ＜ポーズ＞ should be considered stronger than the other.

文の終りのポーズと文中のポーズの前の単語での声の高低の変化に注目して下さい。これは一例に過ぎませんが、このように抑揚を付けて読むことによってあなたの英語がより伝わりやすいものになります。

ほぼ¥完¥璧だと思われますが、いくつかの文でリズムが正確ではない可¥能¥性があります。英語では、前置詞、冠詞、助動詞などの¥機¥能¥語と呼ばれる語は弱く短く発音され、動詞、形容詞、副詞などの内容語と呼ばれる語は強く長く発音される傾向があります。内容語のアクセントのない音節は¥機¥能¥語と同様、弱く短く発音されます。以下はテキストに出てきた文です。

<table>
<tr><td
>the</td><td>North</td><td>Wind</td><td>was</td><td>ob</td><td
>liged</td><td>to</td><td>con</td><td>fess</td><td>that</td><t
d>the</td><td>Sun</td><td>was</td><td>the</td><td>strong</td><
td>er</td><td>of</td><td>the</td><td>two</td></tr>
<tr><td align="center"> 弱 </td><td align="center"> 強 </td><td
align="center"> 強 </td><td align="center"> 弱 </td><td
align="center"> 弱 </td><td align="center"> 強 </td><td
align="center"> 弱 </td><td align="center"> 弱 </td><td
align="center"> 強 </td><td align="center"> 弱 </td><td
align="center"> 弱 </td><td align="center"> 強 </td><td
align="center"> 弱 </td><td align="center"> 弱 </td><td
align="center"> 強 </td><td align="center"> 弱 </td><td
align="center"> 弱 </td><td align="center"> 弱 </td><td

```
align="center">強</td</tr>
</table>

<p>「強」となっている音節は強く長く発音され、「弱」となっている音節は弱く短く発音されます。日本語のようにすべての音節を同じ長さで発音すると英語としてはかなり伝わりにくいものになってしまいます。
</p>

<br><br>
<p>以上でテストは終了です。ご協力ありがとうございました。</p>

</div>
</BODY></HTML>
END_OF_HTML

} elsif ($min == $cate2){
   print"Content-type: text/html; charset=Shift_JIS¥n¥n";
print <<END_OF_HTML;

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd">
<html lang=ja>
<head>
<title>音読自動評価 | Evaluation</title>
<STYLE type="text/css">
```

```
<!--
#cont                                          {line-height:1em;
padding:0.5em;font-size:13px;font-family:Trebuchet
MS;line-height:17px;padding:0.5em 5em; color:#555555}

A:link      { color:#00FFFF; }
      A:visited { color:#6495ED; }
      A:hover     { color:#F5FFFA; }
      a { text-decoration: none }
--></style>

<SCRIPT TYPE="text/javascript">
<!--
var sound2Embed = null;
function sound2Play() {
  if ( !sound2Embed ) {
    sound2Embed = document.createElement("embed");
    sound2Embed.setAttribute("src", "machinegun.wav");
    sound2Embed.setAttribute("hidden", true);
    sound2Embed.setAttribute("autostart", true);
  } else sound2Stop();
  sound2Embed.removed = false;
  document.body.appendChild(sound2Embed);
}
```

```
// -->
</script>


</head>

<BODY bgcolor=#FFF9F9>
<div id="cont">
<h3>evaluation</h3>

<p>このテストでは A, B, C の 3 段階で英語の音読を評価しています。以下があなたの判定結果です。</p>
<p><b>Category B</b>: 母語の影響がしばしばあり、間違った発音をすることもあるが、コミュニケーションを阻害するほどではない。</p>

<h3>単語の発音について</h3>
<p>ひとつひとつの音の発音が正確ではない可¥能¥性があります。例えば、テキストに出てきた last、sun、at の母音を日本語では同じ「あ」と認識しますが、英語では、これらの母音はすべて異なります。ひとつの母音が異なるだけで意味が異なる場合があります。例えば、but と bat は日本語風に発音すると「バット」になってしまいますが、実際には異なる発音です。単語を調べるときは意味や使い方だけでなく、その発音も調べましょう。</p>

<h3>文の読み方について</h3>
<p>文の強勢が正確ではない可¥能¥性があります。単語にアクセント（強勢）がある
```

ように、文にも最も強く発音される単語があります。これは前後関係や話者の意図によってことなりますが、例えば、テキストに出てきた"and at last the North Wind gave up the attempt"という文の場合、一般的に最後の単語 attempt が最も強く発音されます。
イントネーション（抑揚）が不適切な可¥能¥性があります。イントネーションの付け方は一通りに決まるものではありませんが、例えば、文の途中でポーズを置く際に文の終りと同じように声の高低を変化させることは適切ではありません。次の例を聞いて下さい。

例

ここではテキストに出てきた文を次のように 3 つに区切って読んでいます。

They agreed ＜ポーズ＞ that one who succeeded in making the traveler take his cloak off ＜ポーズ＞ should be considered stronger than the other.

文の終りのポーズと文中のポーズの前の単語での声の高低の変化に注目して下さい。これは一例に過ぎませんが、このように抑揚を付けて読むことによってあなたの英語がより伝わりやすいものになります。

以上でテストは終了です。ご協力ありがとうございました。

```
</div>
</BODY></HTML>
END_OF_HTML

} elsif ($min == $cate3){
   print"Content-type: text/html; charset=Shift_JIS¥n¥n";
print <<END_OF_HTML;

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd">
<html lang=ja>
<head>


<title>音読自動評価 | Evaluation</title>

<STYLE type="text/css">
<!--


#cont                                          {line-height:1em;
padding:0.5em;font-size:13px;font-family:Trebuchet
MS;line-height:17px;padding:0.5em 5em; color:#555555}

A:link      { color:#00FFFF; }
```

```
      A:visited { color:#6495ED; }
      A:hover     { color:#F5FFFA; }
      a { text-decoration: none }

--></style>

</head>

<BODY bgcolor=#FFF9F9>
<div id="cont">
<h3>evaluation</h3>

<p>このテストではA, B, C の 3 段階で英語の音読を評価しています。以下があなたの判定結果です。</p>
<p><b>Category C</b>: 発音には母語の影響が強く、聞き手は何を言っているかしばしば聞き返さなくてはならないく、限られた単語、フレーズでも英語の母語話者は注意深く聞かなければならない。</p>

<h3>単語の発音について</h3>
<p>単語に正確な発音にはない音を加えて発音している可¥能¥性があります。例えば、テキストに出てきたwindを日本語風に発音して最後に母音を足したりしてはいませんか？wind の正確な発音は/d/で終わり、その後ろに母音などを加えて発音すると正確に伝わらない場合があります。</p>
<p>いくつかの単語でアクセントを間違えている可¥能¥性があります。例えば、テキ
```

ストに出てきたimmediatelyは2番目の音節が単語の中で最も強く長く発音されます。単語を調べるときは意味や使い方だけでなく、その発音も調べましょう。

<h3>文の読み方について</h3>
ポーズの位置が正確でない可¥能¥性があります。英語を話すとき、読むときはその意味のまとまりごとに区切ってポーズを置きます。カンマやセミコロン、ピリオドでポーズを置くのはそこで意味が区切れているからです。例えば、テキストに出てくる文 "and at last the north wind gave up the attempt"をどこかで 1 回区切るとした場合、"and at last"で 1 回ポーズを置き、"the north wind gave up the attempt"と続けるのが意味が通る読み方です。

以上でテストは終了です。ご協力ありがとうございました。

</BODY></HTML>
END_OF_HTML

}
}
else {
print"Content-type: text/html; charset=Shift_JIS¥n¥n";
print <<END_OF_HTML;

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd">
<html lang=ja>


<head>


<title>音読自動評価 | Instruction</title>

<STYLE type="text/css">
<!--


#cont                                          {line-height:1em;
padding:0.5em;font-size:13px;font-family:Trebuchet
MS;line-height:17px;padding:0.5em 5em; color:#555555}

A:link      { color:#00FFFF; }
      A:visited { color:#6495ED; }
      A:hover     { color:#F5FFFA; }
      a { text-decoration: none }

--></style>
```

```
</head>

<BODY bgcolor=#FFF9F9>
<div id="cont">
<h3>Caution!</h3>
<p>前の録音がアップロードされていません。以下のボタンを押して、前のページに戻って下さい。録音後、"Send"ボタンを押し、ファイルがアップロードされていることを確認して下さい。</p>
<form method="post" action="rec5.cgi">
<INPUT type="hidden" name="filename" value="$filename" >
<br><br>
<p><input type="submit" value="Back"></p>

</form>

</div>
</BODY></HTML>
END_OF_HTML

}
exit (0);

instruction.cgi
```

```
#!/usr/bin/perl -w

use strict;
use CGI;

my ($q, $data, $filename, $name, $cookie, $rest1, $name1, $name2,
$rest2, $rest4, $rest5, $namedef);
my ($q1, $q2, $q3, $q4, $q5, $q6, $q7, $q8, $new, @old, @write);
$q = new CGI;

$q1 = $q->param('q1');
$q2 = $q->param('q2');
$q3 = $q->param('q3');
$q4 = $q->param('q4');
$q5 = $q->param('q5');
$q6 = $q->param('q6');
$q7 = $q->param('q7');
$q8 = $q->param('q8');

$data = $ENV{'QUERY_STRING'};
$data =~ tr/+/=/;
$data =~ tr/&/=/;
($rest1, $name1, $name2, $rest2, $rest4, $rest5) = split(/=/,
$data);
$namedef = $name1 ." ".$name2;
```

```
$name = $name1 ."_".$name2;
$filename = $name;
$filename =~ tr/[A-Z]/[a-z]/;

my ($archive) = "./ques.dat";
$new = "$filename, $q1, $q2, $q3, $q4, $q5, $q6, $q7, $q8¥n";

open(IN, "$archive") or die "Cannot open $archive file!¥n";
   @old = <IN>;
open(OUT, ">$archive") or die "Cannot open $archive file!¥n";
   @write = ($new,@old);
   print OUT @write;
close (OUT);
close (IN);

print"Content-type: text/html; charset=Shift_JIS¥n¥n";
print << "END_OF_HTML";
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd">
<html lang=ja>

<head>

<title>音読自動評価 | Instruction</title>
```

```
<STYLE type="text/css">
<!--
#cont                                                {line-height:1em;
padding:0.5em;font-size:13px;font-family:Trebuchet
MS;line-height:17px;padding:0.5em 5em; color:#555555}

A:link      { color:#00FFFF; }
      A:visited { color:#6495ED; }
      A:hover     { color:#F5FFFA; }
      a { text-decoration: none }

--></style>
<SCRIPT TYPE="text/javascript">
<!--

// -->
</script>
</head>

<BODY bgcolor=#FFF9F9>
<div id="cont">
<h3>instruction</h3>

<p>こんにちは、$namedef さん。</p>
<p>ここで行うテストでは以下のものが必要です。</p>
```

```
<ol>
<li>ヘッドフォン型 PC 用マイクロフォン</li>
<li>Java Runtime Environment</li>
<li>約 10 分間静かで邪魔されない空間</li>
</ol>

<p>Java Runtime Environment をインストールしていない方は<a href="http://java.com/ja/" target="_blank">こちら</a>からインストールして下さい。</p>
<hr>
<p>ここでは以下の文章を文ごとに分けて、音読し、録音します。ここでよく読んで内容を理解して下さい。</p>
<p>
    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 the other.
<br>    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.
<br>    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.
</p>
<p>発音が分からない単語がある場合は事前に調べて下さい。</p>
```

<h3>日本語訳</h3>

北風と太陽がどちらが強いのか言い合っていました。そこに暖かそうなコートを着た旅人が通りかかり、二人は先にその旅人のコートを脱がせた方が強いということに決めました。

北風は力一杯風を吹きつけましたが、強く吹けば吹くほど旅人はコートをしっかりまとい、北風はとうとう諦めてしまいました。

そこで太陽はさんさんと輝き始めました。するとすぐに旅人はコートを脱ぎ、北風は太陽の方が強いと言わなければならなかったのです。

<hr>

それではマイクロフォンの音量をチェックするため、また、このテストの形式に慣れるため、なにか録音してみて下さい。

 以 下 に "Runnning in Free Mode" と 表 ¥ 示 さ れ た ら <img src="img/next.jpg" alt="next"> を押して下さい。

<img src="img/rec.jpg" alt="rec"> を押すと録音が始ります。

<img src="img/volume.jpg" alt="volume"> は音量を示しています。緑が点灯するぐらいの音量で録音して下さい。

録音を終了するには <img src="img/stop.jpg" alt="stop"> を押します。

録音したものを聴く場合には <img src="img/play.jpg" alt="play"> を押します。

録 音 を や り 直 し た い 場 合 に は も う 一 度 <img src="img/rec.jpg" alt="rec"> を押して録音します。

録音が終了したら、<img src="img/send.jpg" alt="send"> を押して提

```
出して下さい。</li>
<li>左下に"SUCCESS"と表¥示されれば完了です。</li>
</ul>
<center><applet

code="com.softsynth.javasonics.recplay.RecorderUploadApplet"
    codebase="../listenup/codebase"
    archive="JavaSonicsListenUp.jar,OggXiphSpeexJS.jar"
    name="ListenUpRecorder"
    width="350"
    height="140">

    <!-- URL for the script which receives the uploaded sound file.
-->
    <param name="uploadURL" value="testrec.cgi">

    <!-- Name of uploaded sound file. Server can change it if needed.
-->
    <param name="uploadFileName" value="whatever.wav">

    <param name="frameRate" value="16000">
    <param name="format" value="s16">

    <param name="compressorEnable" value="yes">
```

```
</applet></center>
<br><br>
<hr>

<p>テストの形式が分かったら、以下の Start ボタンを押してテストを開始して下さい。</p>

<form method="post" action="rec1.cgi">

<INPUT type="hidden" name="filename" value="$filename" >
<p><input type="submit" value="Start"></p>

</form>

</div>
</BODY></HTML>
END_OF_HTML

exit (0);
```

recx.cgi

```
#!/usr/bin/perl -w

use strict;
```

```
use CGI;

my ($q, $filename, $dbfile, @old, @new);

$q = new CGI;

$filename = $q->param('filename');

$dbfile = "./data.dat";

open (IN, "$dbfile") or die "Cannot open $dbfile file!¥n";

@old = <IN>;

open (OUT, ">$dbfile") or die "Cannot open $dbfile file!¥n";
@new = "$filename" . "," . "@old";
print OUT @new;

close (OUT);
close (IN);

print"Content-type: text/html; charset=Shift_JIS¥n¥n";
print <<END_OF_HTML;

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
```

```
"http://www.w3.org/TR/html4/strict.dtd">
<html lang=ja>


<head>


<title>音読自動評価 | Recording</title>

<STYLE type="text/css">
<!--


#cont                                          {line-height:1em;
padding:0.5em;font-size:13px;font-family:Trebuchet
MS;line-height:17px;padding:0.5em 5em; color:#555555}

A:link      { color:#00FFFF; }
      A:visited { color:#6495ED; }
      A:hover     { color:#F5FFFA; }
      a { text-decoration: none }

--></style>

</head>
```

```
<BODY bgcolor=#FFF9F9>
<div id="cont">
<h3>recording 1/5</h3>
<p>以下の文章を読み、録音し、"Send"ボタンを押して提出して下さい。読み間違いがあったり、テキストにない単語を挿入した場合は正確に判定できない場合があります。その際は録音し直して下さい。</p>
<h3>録音する文</h3>
<p>The North Wind and the Sun were disputing which was the stronger when a traveler came along wrapped in a warm cloak.</p><br><br>
<center><applet

code="com.softsynth.javasonics.recplay.RecorderUploadApplet"
    codebase="../listenup/codebase"
    archive="JavaSonicsListenUp.jar,OggXiphSpeexJS.jar"
    name="ListenUpRecorder"
    width="350"
    height="140">

    <!-- URL for the script which receives the uploaded sound file.
-->
    <param name="uploadURL" value="upload_1.cgi">

    <!-- Name of uploaded sound file. Server can change it if needed.
-->
```

```
    <param name="uploadFileName" value="whatever.wav">

    <param name="frameRate" value="16000">
    <param name="format" value="s16">
    <param name="compressorEnable" value="yes">


</applet></center>
<br><br>

<form method="post" action="rec2.cgi">

<INPUT type="hidden" name="filename" value="$filename" >
<p>録音したものを送信したら、下の"Next"ボタンを押して次のページに行って下さい。</p>
<p>"Send"ボタンを押して録音したものを送信しましたか？もう一度確認して下さい。</p>
<br><br>
<p><input type="submit" value="Next"></p>

</form>

</div>
</BODY></HTML>
END_OF_HTML
```

```
exit (0);
```

upload_x.cgi

```
#!/usr/bin/perl -w

use strict;
use CGI;

my ($q, $filename, $dbfile, @old, @new);

$q = new CGI;

$filename = $q->param('filename');

$dbfile = "./data.dat";

open (IN, "$dbfile") or die "Cannot open $dbfile file!¥n";

@old = <IN>;

open (OUT, ">$dbfile") or die "Cannot open $dbfile file!¥n";
@new = "$filename" . "," . "@old";
print OUT @new;
```

```
close (OUT);
close (IN);

print"Content-type: text/html; charset=Shift_JIS¥n¥n";
print <<END_OF_HTML;

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd">
<html lang=ja>


<head>


<title>音読自動評価 | Instruction</title>

<STYLE type="text/css">
<!--


#cont {line-height:1em;
padding:0.5em;font-size:13px;font-family:Trebuchet
MS;line-height:17px;padding:0.5em 5em; color:#555555}

A:link { color:#00FFFF; }
```

```
      A:visited { color:#6495ED; }
      A:hover     { color:#F5FFFA; }
      a { text-decoration: none }

--></style>

</head>

<BODY bgcolor=#FFF9F9>
<div id="cont">
<h3>recording 1/5</h3>
<p>以下の文章を読み、録音し、"Send"ボタンを押して提出して下さい。読み間違いがあったり、テキストにない単語を挿入した場合は正確に判定できない場合があります。その際は録音し直して下さい。</p>
<h3>録音する文</h3>
<p>The North Wind and the Sun were disputing which was the stronger when a traveler came along wrapped in a warm cloak.</p><br><br>
<center><applet

code="com.softsynth.javasonics.recplay.RecorderUploadApplet"
    codebase="../listenup/codebase"
    archive="JavaSonicsListenUp.jar,OggXiphSpeexJS.jar"
    name="ListenUpRecorder"
    width="350"
    height="140">
```

```
    <!-- URL for the script which receives the uploaded sound file.
-->
    <param name="uploadURL" value="upload_1.cgi">

    <!-- Name of uploaded sound file. Server can change it if needed.
-->
    <param name="uploadFileName" value="whatever.wav">

    <param name="frameRate" value="16000">
    <param name="format" value="s16">
    <param name="compressorEnable" value="yes">


</applet></center>
<br><br>

<form method="post" action="rec2.cgi">

<INPUT type="hidden" name="filename" value="$filename" >
<p>録音したものを送信したら、下の"Next"ボタンを押して次のページに行って下
さい。</p>
<p>"Send"ボタンを押して録音したものを送信しましたか？もう一度確認して下さ
い。</p>
<br><br>
```

```
<p><input type="submit" value="Next"></p>


</form>


</div>

</BODY></HTML>

END_OF_HTML


exit (0);
```

Appendix I: The evaluation score given by the three human raters and the three scoring methods

	E01	E02	E03	E04	E05	E06	E07	E08	E09	E10
N	C	A	B	B	A	C	A	C	C	B
K	B	A	A	A	A	A	A	C	C	A
M	C	B	C	C	B	C	B	C	C	C
R1	B	A	C	C	A	B	B	C	C	B
R2	A	A	B	B	A	B	B	C	C	B
R3	A	A	C	C	B	B	B	C	C	B

	E11	E12	E13	E14	E15	E16	E17	E18	E19	E20	E21
N	C	A	B	C	A	A	A	B	C	C	A
K	B	A	B	A	A	A	A	B	A	C	A
M	C	A	C	C	B	A	B	C	C	C	A
R1	C	A	A	C	A	A	A	C	C	C	A
R2	C	A	B	C	A	A	B	C	B	C	A
R3	C	B	B	C	A	A	B	C	B	C	A

Appendix J: The self-evaluation in reading aloud

以下の文章を音読する際にあなたができること、できないことについて自分で評価して下さい。

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.

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.

評価は以下の 6 段階です。1 がもっとも低い評価、6 がもっとも高い評価です。

1. 「まったくできない」
2. 「ほとんどできない」
3. 「少しできない」
4. 「少しできる」
5. 「ほとんどできる」
6. 「十分にできる」

1. 各単語を正しい発音で読むことができる。
2. 各単語内で強く発音すべきところは強く読むことができる。
3. いくつかの単語が連なってひとつの意味になるときは、かたまりとして読み、その前後にポーズを置くことができる。

4. 各文の中で（意味的に）重要な単語は強く読むことができる。
5. 文章の意味を考えて、英語らしい抑揚（イントネーション）をつけて読むことができる。
6. 言いよどむことなく、すらすらと（なめらかに）読むことができる。
7. 意味を考えながら読むことができる。
8. 文法を考えながら読むことができる。