Using Speech Recognition for an Automated Test of Spoken Japanese

Masanori Suzuki
Ordinate Corporation
800 El Camino Real Suite 400
Menlo Park, CA 94015, USA
msuzuki@ordinate.com

Yasunari Harada
Institute for Digital Enhancement of
Cognitive Development
Waseda University
1-6-1 Nishi-Waseda, Shinjuku-ku,
Tokyo 169-8050, Japan
harada@waseda.jp

Abstract

Various kinds of IT and computer technology have enabled language tests to be delivered online or on computer, and to have much faster scoring time. Such computer-based tests can well assess three skill areas such as reading, writing, and listening comprehension. However, in many cases, speaking ability is still inferred by the scores obtained for those three skill areas. Ordinate Corporation and the Institute for Digital Enhancement of COgnitive DEvelopment (the Institute for DECODE) at Waseda University have been jointly working to develop a completely automated test of spoken Japanese, Spoken Japanese Test (SJT)[1]. SJT is intended to provide automated test administration and scoring service by delivering the test over the telephone and by scoring the test using speech recognition technology and other computerized systems. Ordinate and the Institute for DECODE are currently collecting speech data from native and non-native speakers of Japanese to develop a speech recognizer optimized for the Japanese language and for non-native speakers of Japanese.

1. Introduction

Multimedia technology, IT, computers, and voice recognition systems have started to be widely used in foreign language education as instructional aids and as self-learning tools. Within foreign language education, language testing has begun to use these types of technology as well. The most common way to utilize technology in language testing is to develop Computer-Based Tests (CBT). One of the advantages of CBT is to be able to collect test-takers' responses almost immediately and to return scores to the test-takers in a shorter period of time than ever before. Most CBTs assess learners' Listening, Reading, and Writing skills, but they rarely test Speaking. It is still very challenging to test speaking remotely via technology. Therefore, the common format of testing one's speaking ability is still a face-to-face interview format.

More and more foreign nationals are coming to Japan for academic and business purposes. The Japan Foundation (2005) reported that the Japanese Language Proficiency Test was administered to more than 300,000 test-takers in 2004. As more people learn Japanese or as more people come to Japan to study or to work, demand for speaking tests that can measure Japanese learners' speaking skills quickly and reliably is growing. However, few speaking tests are available in the field of teaching Japanese, and, not to mention, there is no speaking test that has utilized computer technology to efficiently test learners' speaking ability.

Ordinate Corporation has developed a language testing system using speech recognition and computerized scoring systems that can evaluate the spoken language skills of non-native speakers. The Spoken English Test (SET) and Spoken Spanish Test (SST) are currently delivered on this

[1] The initial development process was reported at PACLIC 18 held in December 2004 at Waseda University in Japan.

testing system. Both of these tests take about 10-15 minutes to complete and show very highly reliabilities.[2] Ordinate Corporation and the Institute of DECODE (Digital Enhancement of COgnitive DEvelopment) at Waseda University in Japan are currently developing a spoken Japanese test (SJT), which will be administered and scored completely automatically on the Ordinate testing system. A prototype version of SJT has been developed and is currently used for data collection to develop a speech recognizer optimized for the Japanese language.

2. Ordinate Testing System

2.1. Overview

The prototype version of SJT follows the same test administration procedures as other Ordinate tests. An SJT test is administered over a telephone by Ordinate's testing system. Prior to taking a test, a test-taker receives a test paper. One side of it has general test instructions and the other side has a unique Test Identification Number, a telephone number, the verbatim spoken instructions, and examples of tasks and items. When the test-taker calls the telephone number on the test paper, the call gets connected to a TDS (Test Delivery System), then the instructions and first item prompt are pulled from the database, passed over the Virtual Private Network (VPN), and played over the phone, which communicates with the Ordinate testing system via the Internet. In some countries such as Japan, Taiwan, and Korea, a local TDS is set up and test-takers in those countries can take tests using a local toll-free number. The system prompts the test-taker to enter the Test Identification Number printed on the test paper using the telephone keypad. The system then presents the test-taker with a series of spoken prompts in Japanese and the test-taker responds by speaking in Japanese. The system then listens for the test taker's response and after a predefined amount of silence, automatically proceeds to the next item. The system also detects if the test taker hung up the phone before the test was completed. The current prototype version of SJT takes about 15 minutes to complete. Test-takers' responses are then sent from the TDS to the Ordinate testing system via the Internet.

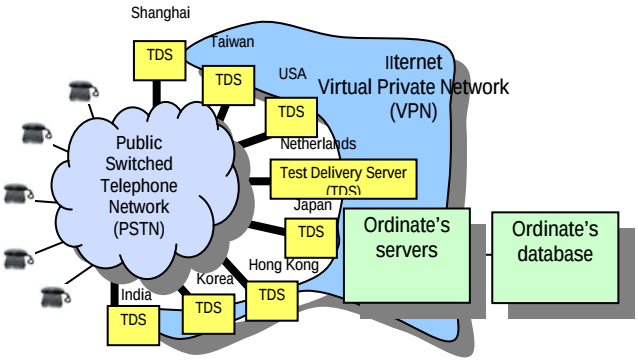

Figure 1. Ordinate Testing System

Upon completion of a test, Ordinate's testing system evaluates the test-taker's responses, generates a score report, and posts it on Ordinate's website, usually within a few minutes. The Ordinate scoring system is generally designed to evaluate two aspects of speech samples: content and manner-of-speaking. The content of the response is scored according to whether the test taker used expected words in the correct sequence. The manner-of-speaking aspect is calculated by measuring the latency of the response, the rate of speaking, the position and length of pauses, the stress and segmental forms of the words, and the pronunciation of the segments in the words within

[2] SET-10 (2003) and SST (2004) have a reliability coefficient of 0.97 and 0.96 respectively.

their lexical and phrasal context. These measures are scaled according to native and non-native distributions and then combined so that they optimally predict human judgments.

A score report consists of one Overall score and four subscores: Sentence Mastery, Vocabulary, Fluency, and Pronunciation. Sentence Mastery and Vocabulary subscores constitute the 'content' aspect, while Fluency and Pronunciation subscores are the 'manner-of-speaking' aspect that the Ordinate scoring system measures. Each score is reported on a scale from 20 to 80. However, the prototype version of SJT is not automatically scored yet. The data collection we are conducting now will enable us to develop an automated scoring system specifically designed for this spoken Japanese test. Figure 1 provides a high-level overview of the testing system.

2.2. Speech Recognition System

As the system receives responses from the test taker, it starts extracting information from the utterance for scoring. The computerized testing system uses an HMM-based automatic speech recognition (ASR), speech to text alignment, and non-linear models to perform automatic scoring.
Each incoming response is first recognized automatically by the speech recognizer that has been optimized for non-native speech. The words, the pauses, the syllables, the phones, and even some subphonemic events are located in the recorded signal.

The speech recognizer itself has several components. One component is an acoustic model. This is a representation of the sounds or *phones* produced when speaking the target language. Two acoustic models need to be built: one for native speakers and one for non-native speakers. In order to develop acoustic models, a significant number of data have to be collected from both native speakers and non-native speakers of the target language. Then, a significant portion of the native and non-native data will be transcribed by native speakers of the target language. Transcribers use Ordinate's Internet-based transcriber interface. This Internet-based transcriber interface allows transcribers to be anywhere in the world and to transcribe anytime they want. The interface allows transcribers to download speech files directly from Ordinate's database.

Another component is a pronunciation dictionary. The pronunciation dictionary lists the most common pronunciations for each word that the system should recognize. A third component is the language model. This is a representation of the sequence of words the speaker is expected to say. For example, if the respondent is asked to repeat the sentence, "It's supposed to rain tomorrow, isn't it," then it is very likely that the test taker will say the words "It's supposed to rain tomorrow, isn't it." The high probability associated with this string of words is encoded in the language model for this sentence. The language models not only contain the most likely strings of words that test takers are expected to say, but also the types of mistakes and disfluencies that non-native speakers are most likely to make. These mistakes and disfluecies get encoded in the language model based on the data collected during data collection.

Using the acoustic models, pronunciation dictionary, and language models, the speech recognition system uses statistical methods to identify the string of words that best matches the respondent's speech. The hypothesis of what the respondent said is then compared to the words in the item. The 'correctness' of the content of the response in addition to the item's difficulty level contribute to the two of the four subscores, that is, Sentence Mastery and Vocabulary.

Other information is also extracted from the respondent's utterance such as speaking time, rate of speech, mean pause duration, and the pronunciation of the segments in the words within their lexical and phrasal context. These and other paralinguistic parameters are then input into non-linear models that are optimized to predict how humans judge the responses with regard to pronunciation and fluency.

Scoring algorithms then process the information from the test and compute scores on the 20-80 test scale. The scores are posted to a secure web site a few minutes after the test is completed and

are made available to those who are authorized to view them such as test takers or test administrators.

As mentioned before, currently, the speech recognition system and computerized scoring system are developed for English and Spanish. In order to develop a speech recognition system and computerized scoring system as described above specifically for Japanese, a significant amount of data both from native and non-native speakers of Japanese needs to be collected, and a significant portion of the collected data will then need to be transcribed by native speakers of Japanese. Currently, data collection for this purpose is in progress mainly in Japan.

3. SJT Construct

The SJT is intended to measure *Facility* in spoken Japanese, that is, the ability to understand spoken Japanese on everyday topics and to respond intelligibly at a native-like conversational pace. All SJT test items are integrated 'listen-then-speak' type requiring the test-taker to employ real-time receptive and productive spoken language skills. In other words, the SJT is designed to measure the test taker's control of these core language processing components.

As shown in Table 1, seven tasks have been developed for the prototype version of SJT. The seven tasks are Reading, Repeat Sentences, Opposites, Short Answer Questions, Sentence Builds, Story-Retelling, and Open Questions. After collecting and analyzing the data, the final set of tasks and the final number of items to be presented in each of the tasks will be determined. Note that the items in Part A: Reading are printed on the test paper and furigana (phonetic aids) has been provided for each of the Chinese characters (kanji) used in the items.

Table 1. Task	Number of Items
Part A: Reading	8
Part B: Repeat Sentences	16
Part C: Opposites	8
Part D: Short Answer Questions	16
Part E: Sentence Builds	10
Part F: Story-Retellings	2
Part G: Open Questions	2
Total	*62*

Table 1. Tasks to be presented in the prototype and the number of items presented in each task

4. Item Writing

SJT test items were developed by native Japanese item writers. As described above, the SJT is intended to measure the ability to understand spoken Japanese on everyday topics and to respond intelligibility at a native-like conversational pace. Therefore, the vocabulary and sentence structure used in the SJT reflect common everyday Japanese. Because the SJT test is intended to measure daily conversational ability in spoken Japanese, some items include honorifics to reflect actual language usage.

After item development, the developed items were reviewed by linguists in Japan and teachers of Japanese as a second language both in Japan and in the U.S. These reviewers were asked to examine the items to see if they conform to the conversational Japanese of educated native speakers and that they use natural expressions. The reviewers were also asked to identify any test

items with expressions specific to only certain areas of Japan. Test items were modified based on the reviewers' comments as necessary, and were recorded by different voice talents which were distinctly different from the examiner voice. These voice talents included both genders and were from the Tokyo area, Osaka, Hiroshima, and Okinawa. Their regional accents were allowed.

5. Normative Data Collection

The prototype version of SJT is now available over the telephone; however, the test is not automatically scored at this point. Since Ordinate's testing system uses automated speech recognition and computer algorithms for automatic scoring, a large amount of normative data is necessary to make SJT automatically graded. Data has to be collected from both native speakers and non-native speakers of Japanese. Currently, data collection is taking place mainly at Waseda University in Tokyo. Participants are mostly undergraduate and graduate native students as well as international students studying at Waseda University. As of October 2005, about 250 native data and 50 non-native data have been collected. Further data collections in different parts of Japan as well as different countries such as Korea, China, and the United States are also scheduled in early 2006. The collected native data will provide us with information about whether the items can be answered correctly by natives regardless of their geographical differences. In addition, the data collection will provide us with samples of different dialects and pronunciation styles. Non-native data will be collected from various first language backgrounds, a wide range of proficiency levels, different age groups, and both genders.

As mentioned before, a significant portion of the collected speech samples from native speaker as well as non-native speakers of Japanese will be transcribed by native speakers of Japanese and the transcriptions will be used to build acoustic models, pronunciation dictionaries, and language models that underlie the automatic response recognition and scoring processes. Finally, with the data collected, item difficulty will be calculated for each item using a testing theory called Item Response Theory (IRT).

6. Validation

After norming data have been collected, a set of human raters will be trained to produce consistent ratings for fluency and pronunciation using scoring rubrics developed by Ordinate. These criterion-referenced scores will be used to train the automated scoring system which will optimally predict the human ratings. A series of validation analyses will also be conducted. One of the analyses to be conducted will be calculating test reliability for the Overall score as well as for the subscores. SET-10 and SST have shown reliability coefficients of 0.97 and 0.96 respectively. SJT is expected to achieve a close reliability coefficient to SET-10 and SST.

Another analysis to be carried out is a comparison of the performance of native versus non-native speakers. We expect that native speakers will obtain high scores on the SJT test, while non-native speakers of Japanese will be distributed over a wide range of scores. The data presented in Figure 2 compares the test scores between native speakers of Spanish and non-native speakers of Spanish using the SST. The distribution of the native speakers clearly distinguishes the natives from the non-native sample. Fewer than 5% of the native speakers received a score below 75, while only 10% of the non-native speakers received a score above 75. The results from this analysis suggest that the SST has high discriminatory power among learners of Spanish as a second or foreign language, whereas native speakers obtain near-maximum scores. If this expected distinction between the two known populations holds for SJT as well, it will support the SJT test's validity.

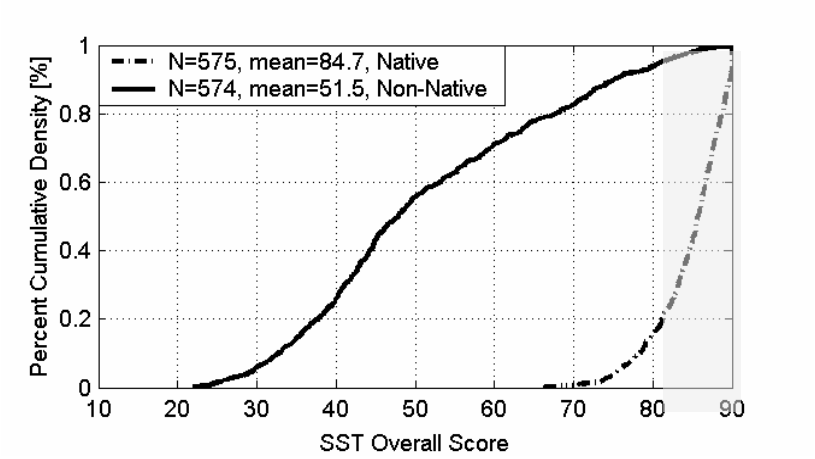

Figure 2. Cumulative distribution functions of SST Overall scores for native and non-native speakers.[3]

Furthermore, a group of human raters will listen to spontaneous responses from story-retellings and open questions, and assign scores to these responses using a set of scoring criteria such as CEFR (Common European Framework of Reference). These estimated scores of test-takers' responses will be compared with their machine-generated scores to see how highly correlated they are.

Finally, we will conduct concurrent validity analyses. We will have a subset of non-native speakers of Japanese take well-established speaking tests such as the ACTFL-OPI (American Council of on the Teaching of Foreign Languages-Oral Proficiency Interview). The purpose of doing this analysis is to understand the relation of SJT scores to the scores obtained from other well-documented human-mediated measures of oral proficiency. For SST, 52 test-takers took both Ordinate's SST and ACTFL-OPI interview tests. As shown in Figure 3, the correlation for these two tests was found to be 0.86, indicating a strong relation between the machine-generated scores and the human-rated interviews. SJT hopes to achieve a strong correlation like SST.

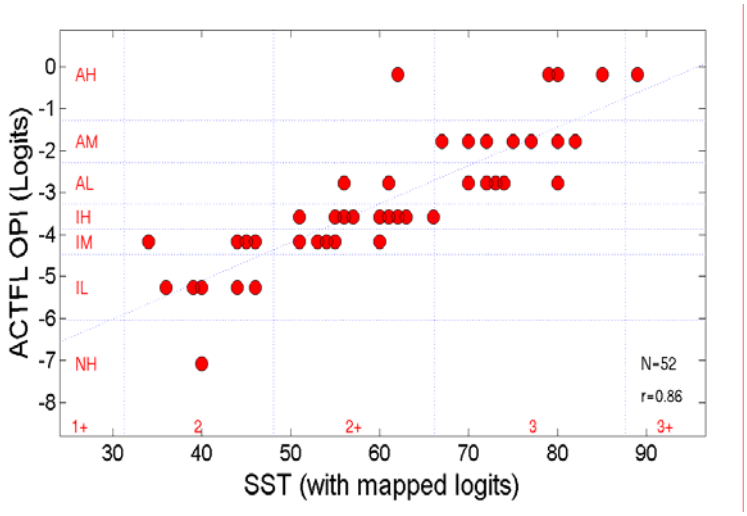

Figure 3. ACTFL OPI scores as a function of SST ratings.

[3] Note that the range of scores displayed in Figure 4 is from 10 through 90, whereas the SST scores are reported on a scale from 20 to 80. Scores below 20 or above 80 are deemed to have saturated the intended measurement range of the test and are reported as 20 or 80 respectively.

7. Summary

Ordinate Corporation and the Institute for DECODE have developed a prototype version of the Spoken Japanese Test. SJT tests Facility in spoken Japanese and will be completely automated. The core element of automatic scoring is a speech recognition engine optimized for non-native speech. The test is currently being used to collect normative data from native and non-native speakers of Japanese. A sufficient amount of collected data as well as transcriptions of the data will enable Ordinate to develop and optimize a speech recognizer for non-native speech patterns. In addition, a series of validation studies will follow the data collection to ensure that the test is reliable and valid. Our hope is that the completely automated SJT will make a significant contribution to the field of teaching and learning Japanese as a second/foreign language.

8. Acknowledgement

The research reported here is partly supported by Grant-in-Aid for Exploratory Research #16652040, provided by the Japanese Ministry of Education, Culture, Sports, Science and Technology, and by funding from Ordinate Corporation.

9. References

Japan Foundation. (2005). Nihongo noryokushiken kekkano gaiyo [Results of the Japanese Language Proficiency Test administered in 2004]. Retrieved on May 31, 2005 from http://www.jees.or.jp/jlpt/pdf/result_2004_3.pdf

Ordinate Corporation. (2004a). SET-10 test descrip-tion & validation summary. Menlo Park, CA: Author.

Ordinate Corporation. (2004b). SST test description & validation summary. Menlo Park, CA: Author.

Harada, Y., Suzuki, M., & Bernstein, J. (2004) Developing an Automated Test of Spoken Japanese. *Proceedings of PACLIC 18: The 18th Pacific Asia Conference on Language, Information and Computation*. 291-298.