ELSEVIER

Speech Communication 30 (2000) 167–177

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

An interactive dialog system for learning Japanese

Farzad Ehsani [a,*], Jared Bernstein [b], Amir Najmi [c]

[a] *Sehda, Inc., 1040 Noel Drive, Menlo Park, California 91025, USA*
[b] *Ordinate Corporation, 1040 Noel Drive, Menlo Park, California 94025, USA*
[c] *Stanford University, Stanford, California 94305, USA*

Received 2 February 1998; received in revised form 4 January 1999; accepted 15 March 1999

Abstract

Subarashii is a system that uses automatic speech recognition (ASR) to offer first-level, computer-based exercises in the Japanese language for beginning high school students. Building the Subarashii system has identified strengths and limitations of ASR technology. The system was tested with 34 students at Silver Creek High School in San Jose, California and with 13 students at Stanford University in Stanford, California. Recognition accuracy was measured and user errors were analyzed. The *functional accuracy* defined as the percentage of time when the system performs the correct functional behavior turned out to be generally higher than the per-utterance speech recognition accuracy.

Keywords: Speech recognition; Computer tutor; Japanese language

1. Introduction

During the past two decades, oral communication skills have received increasing attention (Warschauer, 1996). Students' ability to engage in meaningful conversational interaction in a target language is considered an important, if not the most important goal of second language education. For many people, the sheer number as well as the visual complexity of characters in languages such as Japanese and Chinese make reading and writing those languages especially difficult to learn. When learning an ideographic writing system, it may take more than three years of study to learn enough characters to read a simple newspaper article. For this reason written exercises and reading practice have a diminished role in supporting the development of basic oral proficiency in such ideographic languages.

Live language teachers are essential for teaching oral communication skills, however they usually have to be shared among many students in a class. An increasingly viable alternative to a language instructor which could provide intensive individual conversation is a computer-based Interactive Spoken Language Education (ISLE) system that understands what a student is saying (within a constrained context) and responds in pedagogically meaningful ways. An ISLE system should relieve the teacher of some routine tasks such as engaging beginning students in spoken language production (Warschauer, 1996). To this end, an ISLE system can use recent advances in the field of speech recognition technology for the purposes of Japanese language instruction.

* Corresponding author. Tel.: +1-650-328-8877; fax: +1-650-328-8866.
E-mail address: farzad@sehda.com (F. Ehsani).

0167-6393/00/$ - see front matter
PII: S0167-6393(99)00042-4

The key to the future of multimedia computer-aided language learning (CALL) systems will be their ability to understand and judge continuous spoken language with programmable levels of acceptance (Bernstein et al., 1990; Neumeyer et al., 1996; Waters, 1995). Furthermore, the system should be able to simulate essential features of human–human communication. That is, interactions should work without requiring collateral cues from a mouse or keyboard, they should operate at an appropriate conversational pace and they should incorporate verbal strategies for resolving misunderstandings. While only using simple rejection based on pruning, the Subarashii system explores some aspects of speech recognition and user interface technology that may form the basis of advanced ISLE systems for any language. What we need to know are: (1) can contextually limited but meaningful scenarios be designed and authored in a relatively short time, (2) is the existing recognition technology advanced enough for useful interactions, and (3) are these interactive scenarios effective in producing measurable gains in the learner's language proficiency.

2. System overview

Subarashii, meaning "wonderful" in Japanese, is an interactive spoken language education system for the beginning student of Japanese. The Subarashii system offers learners of Japanese the opportunity to solve simple problems through (virtual) spoken interactions with monolingual Japanese natives. Subarashii is an ISLE system designed to understand what a student is saying in Japanese (within a constrained context) and to respond in a meaningful way in spoken Japanese. The computer system poses problems in written English and offers occasional support to the student in the form of written reminders, but problems can only be solved by speaking and understanding Japanese.

From the vantage point of an instructional designer, the first focus is the selection of content to present in an encounter. Encounters were chosen from typical, everyday situations that a student might encounter – for example, meeting someone for the first time; making, accepting, or refusing an invitation; buying something at a grocery store; or visiting a restaurant. Though the primary focus in Subarashii has been on high school students, we selected situations that are also relevant to adult students of Japanese. Because many forms in the Japanese language depend on the gender and age of the interlocutors, designing such a selection was no simple matter. The encounters deployed in Subarashii are somewhat similar to the progression of material in *Yookoso* by Yasuhiko Tohsaku (1994). See also (Jordan and Nadi, 1987).

The goal of the Subarashii project has been to extend the range of activities available in interactive spoken language education systems, and to demonstrate the effectiveness of these activities. The system evaluation provides preliminary evidence that meaningful conversational practice can be authored and implemented using an ASR system.

Our strategy attempts to predict errors that a student is likely to make. These errors are generally not random, but follow patterns that are recognizable to an experienced language teacher. Our approach has been not to reject utterances on the basis of deviation from a single "gold-standard" model of the correct response. Subarashii compares each utterance both to a model of the correct response and to a set of models of likely incorrect responses. The model (correct or incorrect) that most closely matches the utterance to be recognized will be what the computer understands the speaker to have said. In the event that none of the models compare well with the utterance given, the computer rejects the utterance.

3. System architecture

To understand how the system is constructed, consider Fig. 1 which shows the structure of the program modules on which Subarashii is based.

All modules, except the audio/ASR module, were implemented in the Java programming language on an SGI Indy workstation. The audio/ASR module was implemented on top of HTK (Hidden Markov Model ToolKit) in the C programming language, and communicates with the other modules through Unix sockets. Both the presentation

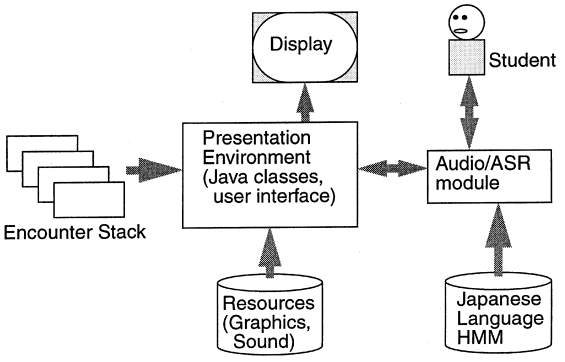

Fig. 1. Subarashii system diagram.

environment and the audio/ASR module were independent of the specifics of the exercises that are part of Subarashii. They simply allow for the implementation of a high-level object called a *card*. This 'card' class encapsulates subroutine calls to display pictures, play sounds, get verbal and graphical input from the student, etc. Each encounter comprises a series of cards called a *stack*, and resources such as sounds, pictures and recognition grammars. Each card represents a single exchange between a virtual interlocutor and the student. The card is designed to be flexible enough to allow for any kind of spoken language exercise. During the course of the encounter, the interaction is implemented by a particular sequence of card presentations from within the stack, depending upon what the student has said.

We used finite state grammars in the Backus Normal Form (described in (Young et al., 1997)) to represent the recognition grammar for each card object. Each card object has its own scene-specific finite state grammar. The author of a particular encounter is expected to create a finite state grammar for each scenario. This grammar enables the author to specify in a compact way how each expected input from the student should be processed. The ASR module uses this recognition grammar together with a set of Hidden Markov Models (HMMs) to recognize what the student says. The recognition grammar is specified in romanized Japanese and must be parsed and translated into a phoneme-level specification (a recognition network) as required by the ASR module. This is achieved automatically by the use of an automatic recognition grammar compiler. The recognition module uses HMMs that have been designed to be speaker independent and to accommodate a wide variety of non-native accents.

We also used finite state grammars to do the dialog management. A number of previous studies have shown that such grammars are adequate for restricted sub-language communication with computers (Dahlback and Jonsson, 1992; Levinson, 1981; Bilange, 1991). We need a prototyping environment to collect user-data to build such grammars. We implemented this by using a traditional Hypercard environment on a Macintosh computer with text input and output. Granted that text responses (right or wrong) will not be identical to students' spontaneous spoken responses, we assumed that they would provide useful information in the development process. Hypercard provides an efficient means of modifying each encounter to accommodate the observed text input from a test group of high school students.

The Hypercard prototyping environment automatically recorded every student input received and organized it in order of frequency. Thus, the author of the encounter can decide at a glance what a student is likely to say at any point, providing an appropriate course of events, including feedback for common errors. Once the logic of an encounter has been verified to the author's satisfaction (perhaps after several prototyping iterations), all requisite system responses to the user input are recorded, recognition networks are generated, and Hypercard scripts are translated into Java. The prototyping environment limits the features of HyperTalk (Hypercard's programming language) that can be used in a prototype encounter to a simple set of commands that are easily translated into Java.

4. Tasks

The current implementation of Subarashii that is described here had four encounters. The encounters, in order of increasing complexity and difficulty, were:

- Glad to Meet You (GTMY).
- Movie Friday? (MF).

- Are You Busy? (AYB).
- Got Milk? (GM).

After selecting an encounter, a starting screen appeared. It had an opening graphic display and a written mission statement. As students experienced the encounters, they progressed from a very passive encounter, "Glad to Meet You", where they only responded to system-initiated prompts, to "Got Milk", where the student had to take all the initiative.

Fig. 2 shows the state grammar diagram of the first three encounters; the fourth was much more complicated, even at the schematic level shown in Fig. 2. Each square represents a turn in the conversation, and each oval represents a set of responses that allow the student to advance to the next dialog turn. Self-loops represent multiple error paths that could be taken, which return the dialog to the same state. Multiple ovals in the same path represent different sets of responses that can be made by the students. Each oval elicited a different response from the system, but all of them led to the same next dialog state.

A more detailed picture of the sub-networks around the first two states in the third encounter can be seen in Fig. 3 with the Japanese utterances translated to their English equivalents. In this figure, ovals represent Japanese spoken by the system, rectangles represent possible Japanese utterances by the student, and shaded rectangles represent written English comments to the student that were displayed, as appropriate, on the screen. As can be seen, even this partial sub-network is quite complicated. The user utterances "Good day" and "It's been a long time, hasn't it?" are responded to with a different initial response, but both responses are followed by "Are you busy this evening?" thereby leading to the next interaction. In the case of user errors such as responding to "Good day" with "Pleased to meet you," instead of giving an accepted answer, the system simply tells the student that the response is inappropriate in this situation, and lets the student try again. An improved version of this system should give the right answer or at least better hints about the right answer after 2 or 3 attempts by the student.

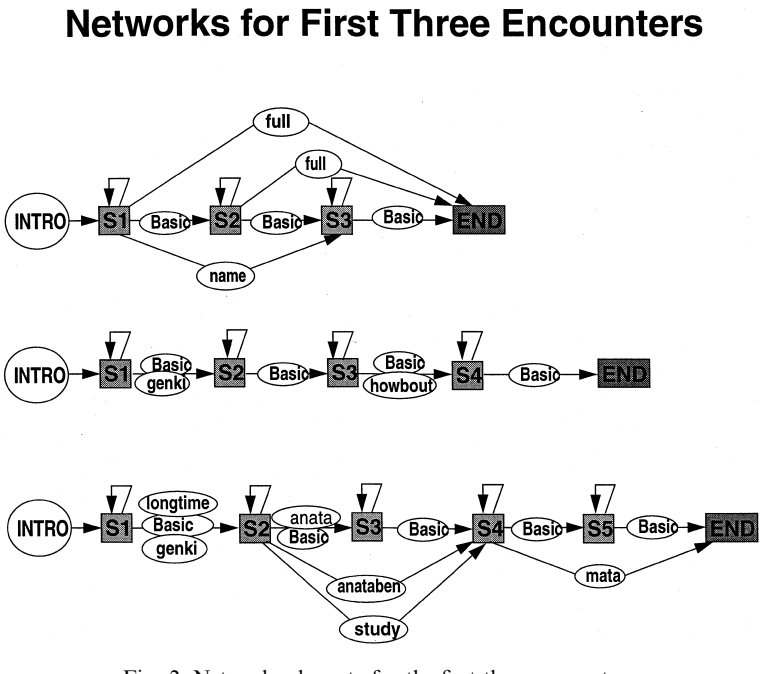

Fig. 2. Network schemata for the first three encounters.

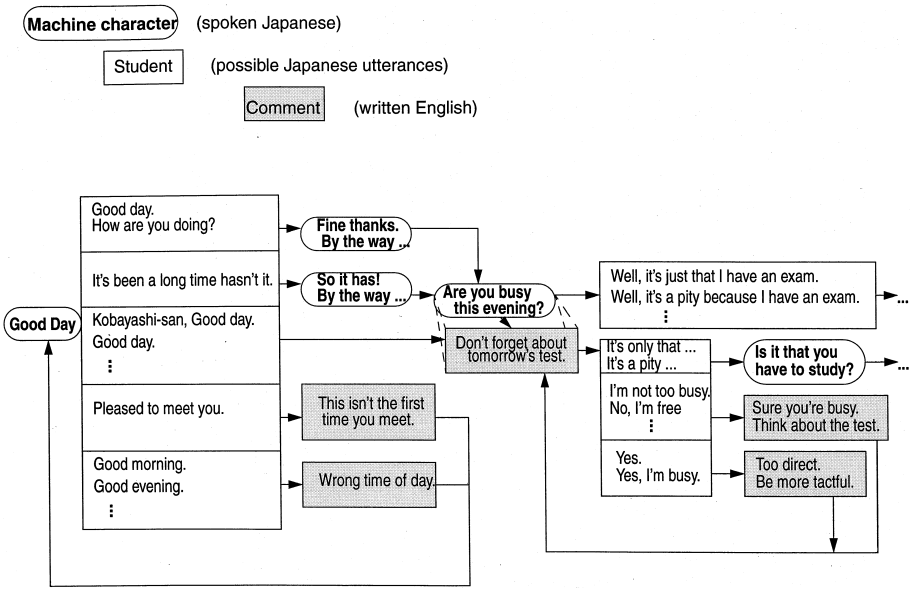

Fig. 3. Initial network for "Are You Busy?".

5. Evaluation

5.1. Data collection and results at Silver Creek

The Subarashii system was set up at Silver Creek High School in San Jose, California. At Silver Creek, 34 students currently enrolled in Japanese language classes went through the encounter. The student sample includes 12 first year students, 11 second year students, 8 third year students, and 3 fourth year students. In the sample of students, most had an A or B grade average in Japanese, although there were a few C and D students.

All of the students were taught by an instructor who was also involved in developing the training material, and a few of the students had done the original Hypercard exercises that were used for prototyping the encounters.

The encounters were preceded by a user survey that asked for the student's class level, their most recent term grade in Japanese, and about their general experience with Japanese. The encounters were followed by a second set of questions about their experience in the encounters and their ability to interact with the system. Four students, two in the first year, and two in the second year, went through the exercises a second time on their own initiative a few days after they had taken the test, initially making a total of 38 sessions. The test was conducted in a quiet classroom after class time. The students were offered cookies and drinks after their participation, but no extra credit was given.

Each student's responses were transcribed, and human graders categorized each response as grammatically and pragmatically correct or incorrect. Each response was judged by at least two of the graders. A certain number of transcribed responses were judged as operator errors that were not rejected by the system. These were silent responses or problems with starting the recognizer which resulted in responses being cut off. Furthermore, each response was judged to be in the grammar network or outside of it. Finally, the recognition accuracy for each response was determined.

Table 1 gives a rough analysis of the student responses for the 38 sessions. Specifically, it shows the percentage of *in-network*, *out-of-network* and *operator-error* phrases. The first two categories are further subdivided into *correct grammar* usage or *incorrect grammar* usage by the student. Finally,

Table 1
User and recognition behavior for Silver Creek students percent response by category

	In-network		Out-of-network		Operator
Grammar	Correct	Incorrect	Correct	Incorrect	Error
GTMY	46.9 (92.7)	0	1.7	45.7	5.7
MF	50.2 (58.4)	2.8	7.6	33.7	5.6
AYB	63.9 (87.1)	8.2	13.9	10.3	3.6
GM	54.2 (67.4)	0	27.7	14.5	3.6
Total	53.7 (74.7)	2.7	13.6	25.4	4.6

the recognition accuracies of the system for *in-network* phrases are shown in parentheses.

On average, the grammatically incorrect *in-network* phrases were only used 2.7% of the time although these phrases constitute about 38% of all the paths in the grammar. Note that "Glad to Meet You" which is the easiest encounter, had the highest recognition accuracy for *in-network* phrases; however, only 46.9% of the responses were *in-network*. Most of the errors in this encounter are due to students using their own name while introducing themselves (as opposed to using their assumed name: Smith). Also note that there does not seem to be a decrease in recognition accuracy with the more difficult encounters such as "Movie Friday", which is somewhat easier than "Are You Busy" or "Got Milk". In fact, we did not see any significant correlation between grade average, number of years studying Japanese or general interest, and recognition accuracy or correct grammar usage. We may assume, therefore, that even the fourth encounter is not far beyond the competence of first year students of Japanese.

We also looked at the number of dialog turns (utterances) each student took to complete each task. Fig. 4 shows a plot of the number of turns (for the "Got Milk?" and the "Movie Friday" encounters) that each student took to complete the encounter versus the number of machine-response states traversed (e.g. S1, S2, etc. in Fig. 2). The size of the symbols indicates the number of students. Most of the sessions are completed by the students with a number of dialog turns only slightly larger than the number of states traversed. This was also the case in the other two exercises. This tells us that, except for a few students who get stuck in a particular state, most of the students seem to go through the interactions very quickly.

Table 2 shows the median, expected value and standard deviation for the number of turns and the number of states traversed by each student. On average, a student takes 1.4 turns to go through each state. These advances consist of correctly

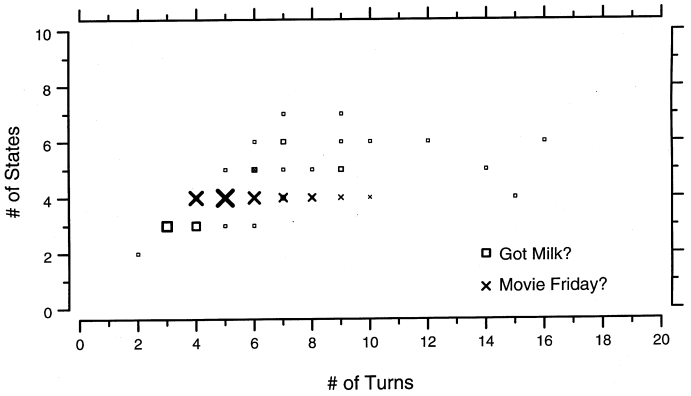

Fig. 4. States traversed versus turns taken.

Table 2
Number of turns and states for Silver Creek students

Grammar	Number of turns			Number of states		
	Median	Expected value	Standard deviation	Median	Expected value	Standard deviation
GTMY	4	4.6	2.3	3	3.0	0.6
MF	6	6.6	4.7	4	4.0	0.7
AYB	5	5.1	1.4	5	4.7	1.0
GM	6	6.6	3.6	4	4.4	1.5
Total	22	22.8	6.9	16	16.1	3.0

recognized correct phrases and incorrectly recognized (as correct) incorrect phrases. Table 1 shows that 53.7% of the utterances have both correct grammar usage and are in the grammar network, although only 74.7% of these utterances are correctly recognized by the system. This means that only 40% of all the phrases are correctly recognized and correct *in-network* sentences. If 70% of the utterances cause an advance to the next dialog state, but only 40% follow designed paths, then it would seem that 30% of the utterances must be incorrectly accepted as true. However when we re-examined the data, this was not the case.

We noticed that the system was often behaving better than the speech recognizer. This discrepancy is caused by the way that we have been looking at the data. A large percentage of utterances that were "mis-recognized" were recognized such that the system responded appropriately and had the same functional behavior that it would have had if the utterance had been correctly recognized. If we re-analyze the data and include every recognition output that produced the right behavior under the correct category, we get a radically different view as shown in Table 3.

Table 3 shows the functional accuracy in each category as well as the combined total. The number in parentheses in the last column indicates the serendipitous percentage increase in functional accuracy as compared with the previous accuracy measurement. In this case, the percentage of correctly recognized correct sentences is increased to 52%, which indicates an 18% false acceptance rate.

5.2. Data collection and results at Stanford University

Based on the *out-of-network* utterances that we found at Silver Creek, we modified the grammar networks to include some of the more common errors that we encountered in our pilot data collection. For example we had included the equivalent Japanese phrase for "Thank you" in the initial Greeting-state in the Glad-to-Meet-You encounter, and we had not included the equivalent Japanese phrase for "Good day". Results in Silver Creek showed that the former is not a common mistake and therefore should not be included in the grammar network, while the latter is common and therefore should be included. However, at the

Table 3
Functional behavior for Silver Creek students percent response by category

Grammar	In-network		Out-of-network		Combined
	Correct	Incorrect	Correct	Incorrect	Total (increase)
GTMY	93.9	0.0	33.3	42.5	64.0 (+20.6)
MF	64.0	100	78.9	60.7	61.4 (+29.3)
AYB	63.9	62.5	88.9	65.0	85.1 (+26.3)
GM	77.0	0.0	46.4	38.9	60.2 (+23.7)
Total	81.3	73.9	61.0	50.9	66.9 (+23.3)

Table 4
User and recognition behavior for Stanford students percent response by category

	In-network		Out-of-network		Operator
Grammar	Correct	Incorrect	Correct	Incorrect	Error
GTMY	60.4 (100)	0 (0)	8.3	27.1	4.2
MF	48.3 (48.8)	0 (0)	30.3	19.1	2.2
AYB	55.4 (91.7)	0 (0)	27.7	10.8	6.2
GM	32.9 (78.6)	1.2 (0)	43.5	17.6	4.7
Total	47.4 (77.2)	0.3 (0)	30.0	18.1	4.2

same time, we expected the results to be worse due to the difference in population, teaching style, as well as Japanese proficiency level. During the summer of 1997, the Subarashii system was set up at Stanford University. Except for the slightly improved networks, we tried to keep the conditions at Stanford similar to Silver Creek. Data collection was done after hours, and in this case the students were offered a small sum of money for their participation. Similar instructions were given to the students for the collection process.

At Stanford, 13 students, enrolled in summer intensive Japanese language classes (first year through third year) went through the encounters. The Stanford sample included three first year students, five second year students, and five third year students. All but one of the students in the Stanford sample had an A or B grade average in Japanese. Unlike at Silver Creek, neither the instructors nor the students at Stanford had ever seen the Subarashii system or the Hypercard exercises with which the encounters had been prototyped. Furthermore, the text book at Stanford was different from the one used at Silver Creek. Finally, most of the students that we talked to were taking Japanese not to fulfill a requirement but because they were very much interested in living in Japan for some period of time.

As at Silver Creek, the Stanford encounters were preceded by a user survey that asked for the students' class level, their most recent term grade in Japanese, and about their general experience with Japanese encounters and their assessment of their ability to interact with the system. The encounters were followed by a second set of questions about their experience in the encounters and the assessment of their ability to interact with the system.

Table 4 gives a rough breakdown for the 13 student sessions at Stanford. Specifically, it shows the percentage of *in-network*, *out-of-network* and *operator-error* phrases. Again, as in Table 1, the latter two are further subdivided into *correct grammar* usage or *incorrect grammar* usage. Finally, the recognition accuracies of the system for *in-network* phrases are shown in parentheses. On average, the grammatically incorrect in-network phrases were used only 0.3% of the time, although they constitute about 38% of all the paths in the grammar. Note that "Glad to Meet You", which is the easiest encounter, had the highest recognition accuracy for *in-network* phrases, however only 60.4% of the responses were actually *in-network*. Again, similar to Silver Creek, most of the errors in this encounter were due to students using their own name while introducing themselves. Also note that there does not seem to be a decrease in recognition accuracy with the more difficult encounters. In fact, "Movie Friday", the second easiest encounter, seems to have the lowest accuracy as with the Silver Creek students. This told us that perhaps vocabulary alone is not the only factor determining success with these encounters. Among the students in the Stanford sample, we did not see any significant correlation between student demographics and either recognition accuracy or correct grammar usage. Finally, note that the total recognition accuracy for the Stanford students seems to be slightly lower than for the Silver Creek students (36.6% as opposed to 43.6%).

Table 5 shows the median, expected value and standard deviation for the number of turns and the number of states traversed by each student. Again, on average, a student takes 1.4 turns to go through each state. These results look very similar to those of the Silver Creek students, with the exception

Table 5
Number of turns and states for Stanford students

Grammar	Number of turns			Number of states		
	Median	Expected value	Standard deviation	Median	Expected value	Standard deviation
GTMY	4	3.7	1.6	3	3.0	1.2
MF	6	6.8	4.3	4	4.2	1.4
AYB	5	5.0	2.1	5	4.5	1.4
GM	4	6.5	5.7	4	4.5	2.3
Total	19	22.1	9.4	16	16.2	5.1

that the median number of turns for "Got Milk" for Stanford students is lower: four responses for Stanford, six responses for Silver Creek.

If we look at the functional accuracy again we get a radically different result from that shown in Table 4. Table 6 shows the functional accuracy in each category as well as the combined total accuracy (similar to Table 3). The number in parentheses in the last column indicates the increase in functional accuracy from the *in-network* correct accuracy measurement shown in the left column of Table 4. Note that these results were as good and in some cases better than the results achieved at Silver Creek High School, even though the raw recognition accuracy for the Stanford students was slightly lower.

5.3. User surveys

We conducted two user surveys, one before and one after the students' experience with Subarashii. Neither survey revealed any significant relation between survey results and performance in the encounters. The post-Subarashii survey solicited degrees of agreement or disagreement with various assertions, such as "The computer understands everything that I say.". In this questionnaire students choose one response among: *strongly agree, agree, neutral, disagree* and *strongly disagree*. For evaluation purposes we assigned a number between 1 through 5 to each one of their responses with 1 representing *strongly disagree* and 5 representing *strongly agree*.

Students at Stanford indicated a slightly lower level of comfort interacting in Japanese (average 3.8) than the Silver Creek students (average 4.1) although they expressed more confidence that they understand the computer (4.9 at Stanford, 4.4 at Silver Creek). They had about the same level of confidence that the computer understood them (4.0 at Stanford, 3.9 at Silver Creek). Stanford students were more neutral about the statement that they are better speakers of Japanese as a result of using the system (3.1 at Stanford, 3.6 at Silver Creek). All the students wanted to use the program again, but the university students were somewhat less enthusiastic than the high schoolers (4.5 at Stanford, 4.7 at Silver Creek). The Stanford group thought that Subarashii could further improve their Japanese if used again, but not to the same degree as the Silver Creek students (3.9 at Stanford, 4.5 at Silver

Table 6
Functional behavior for Stanford students percent response by category

Grammar	In-network		Out-of-network		Combined
	Correct	Incorrect	Correct	Incorrect	Total (increase)
GTMY	100	0	25.0	69.2	81.2 (+20.8)
MF	86.0	0	63.0	64.7	73.0 (+49.4)
AYB	94.4	0	77.8	84.7	83.1 (32.2)
GM	90.4	0	59.3	59.6	55.3 (+29.4)
Total	90.4	0	59.3	59.6	71.4 (+34.8)

Creek). All of these results showed a very positive attitude towards the system, its user-interface and speech recognition.

5.4. Observations on user behavior

Attending the data collection process was a very valuable experience for the authors as we learned about existing and potential problems with our user-interface. Almost all of the students at both Silver Creek and Stanford University had some trouble putting on and getting used to the head-mounted microphones. However, once the students got adjusted to the microphone they seemed to have less trouble putting it on a second time. We noticed that the students had substantial difficulty with the push-and-hold-to-speak button. This kind of interface has been especially useful when doing demonstrations, but many students struggled to get used to it. We also noticed that because we did not have direct volume adjustment in the program, many of the initial errors in the system were due to poor sound quality with the amplitude being either too low or too high. Furthermore, we observed that our reject scheme is not working as well as we hoped, as a lot of false starts and babble were being accepted as valid utterances. Finally, it was clear that the Stanford students, especially the more advanced students, really tested the system's limitations as they often tended to string multiple sentences together.

The extent of the authenticity of our simulations came out during one of the data collection sessions at Stanford. One of the female students going through the ''Movie Friday'' interaction was supposed to ask the animated Japanese male student out to a movie, but blushed deeply and was unable to continue the session.

6. Ongoing work

In 1997–98, Entropic completely rebuilt Subarashii using Macromedia Authorware and Entropic's Application Programming Interface for ISLE applications (Ehsani et al., 1997). The program now runs under the Windows 95 operating system on Pentium 133 with the recognition engine occupying less than 8 Mbytes of memory. We added supporting Japanese exercises which a student can use in stand alone mode to acquire some of the language needed for the encounters and made the user interface easier to navigate. The speech recognition interface was modified from a push-and-hold-to-speak to just push-to-speak – that is, the users push the button once before the start of their utterance and a silence detection mechanism determines the end point. Using data collected in Japan, we trained new acoustical models which have significantly improved the baseline recognition accuracy. We now use confidence scoring to reject utterances which are very unlikely in the recognition networks. We plan to continue experimenting with pronunciation scoring to see what sort of meaningful feedback we can provide for *in-network* utterances. Finally, we plan to re-evaluate the system at a local university or high school in the near future. Meador et al. (1998) provide a description of the recent system modifications and performance data.

7. Conclusion

The current Subarashii system provides preliminary evidence that meaningful conversational practice can be authored and implemented in a relatively short time (~4–5 person-months). We assume that with the availability of commercial and research-oriented dialog building technology such as those from Unisys, OGI, SpeechWorks, etc., developers will be more efficient at building such systems, and the time to author these encounters will be reduced. Also, these encounters seem both useful and enjoyable to students, and current ASR technology can support realistic interactions. Subarashii also showed that ''open-ended'' dialog practice for users with limited language ability and non-native pronunciation is possible. We did not explicitly measure the system's effectiveness in producing measurable gains in the language learner's proficiency, and as yet we do not have enough material to constitute

such a test. However, the users' subjective responses did indicate a strong willingness to re-use the system as well as their belief that Subarashii could further improve their Japanese if used again.

The evaluation of the system also suggests that strict recognition accuracy is in fact not a good yardstick for measuring the effectiveness of the system. Because of the way the interactions were designed, even with a low average utterance correct recognition accuracy of 36.6% and 43.6%, the system was able to achieve reasonable functional accuracy of 71.4% and 66.9%. These encouraging results show that near perfect recognition accuracy may not be a strict requirement for an effective speech dialog system.

Acknowledgements

The Subarashii project was conducted by the Language Systems Group of Entropic, Inc., in Menlo Park, California. The work described in this paper was funded in part by the U.S. Department of Education and by the Federal Language Training Laboratory. The authors gratefully acknowledge essential assistance from Rushton Hurley of Silver Creek High School in San Jose, California as well as Katheleen Egan and Steve Stokowski of Federal Language Training Laboratory.

References

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., Weintraub, M., 1990. Automatic evaluation and training in English pronunciation. In: Proceedings ICSLP-90, Kobe, Japan.

Bilange, E., 1991. A task independent oral dialogue model. In: Fifth Conference of the European Chapter of the Association for Computational Linguistics, Berlin.

Dahlback, N., Jonsson, A., 1992. An empirically based computationally tractable dialogue model. In: Proceedings of the 14th Annual Meeting of Cognitive Science Society, Bloomington, Indiana, pp. 785–790.

Ehsani, F., Bernstein, J., Najmi, A., Todic, O., 1997. Subarashii: Japanese interactive spoken language education. In: European Conference on Speech Communication and Technology, pp. 681–684.

Jordan, E.H., Nadi, M., 1987. Japanese: The Spoken Language, Parts I, II, III. Yale University Press, New Haven, CT.

Levinson, S., 1981. Some pre-observations on the modelling of dialogue. Discourse Processes 4, 93–116.

Meador, J., Ehsani, F., Egan, K., Stokowski, S., 1998. Interactive dialog system for learning Japanese. In: Proceedings of the ESCA Workshop on Speech Technology in Language Learning (STiLL 98), Marholmen, Sweden, 1998, pp. 65–68.

Neumeyer, L. et al., 1996. Automatic text-independent pronunciation scoring of foreign language students. In: International Conference on Spoken Language Processing, pp. 1457–1460.

Tohsaku, Y.-H., 1994. Yookoso! An Invitation to Contemporary Japanese, McGraw-Hill, New York.

Warschauer, M., 1996. Computer-assisted language learning: An introduction. In: Fotos, S. (Ed.), Multimedia Language Teaching, Logos International, Tokyo, pp. 3–20.

Waters, Richard, 1995. The audio interactive tutor. Computer Assisted Language Learning 8 (4), 325–354.

Young, S. et al., 1997. The HTK Book, Entropic Inc.'s Hidden Markov Model Toolkit Manual, pp. 171–179.