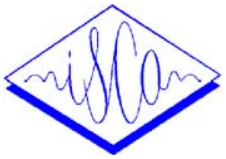

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

SUBARASHII: JAPANESE INTERACTIVE SPOKEN LANGUAGE EDUCATION

Farzad Ehsani, Jared Bernstein, Amir Najmi, Ognjen Todic
Entropic Research Laboratory, Inc.
1040 Noel Dr., Menlo Park, CA 94025, USA
Tel: 1-415-328-8877, FAX: 1-415-328-8866, E-mail: farzad@entropic.com

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 and has led to some novel methods in the development of materials for computer-based interactive spoken language education.

10.21437/Eurospeech.1997-239

1. INTRODUCTION

Speech is the primary medium of a living language. For English-native learners, the complexity and unfamiliarity of the written form of languages such as Japanese make proficiency in the spoken language an especially important part of training and maintenance of language skills. A live language teacher usually has to be shared among many students in a class, but intensive individual conversation could be provided by a computer-based Interactive Spoken Language Education (ISLE) system that understands what a student is saying in Japanese (within a constrained context) and responds in meaningful ways. An ISLE system should relieve the teacher of some routine tasks such as engaging beginning students in spoken language production. To this end, an ISLE system can use recent advances in the field of speech recognition technology for the purposes of Japanese language instruction.

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 [1] [2]. 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 those aspects of speech recognition and user interface technology that will form the basis of advanced ISLE systems for any language. What we need to know is (1) which interactive formats can (or cannot) be supported by high performance speech recognizers running on the coming generation of computers, and (2) which of the feasible interactive formats are most enjoyable for users and most effective in producing measurable gains in language proficiency.

2. SYSTEM OVERVIEW

The Subarashii system offers beginning students 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.

Entropic's 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 compares well with the utterance given, the computer rejects the utterance. This strategy attempts to predict errors that a student is likely to make. These errors are not random, but follow patterns that are recognizable to a skilled language teacher.

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 with an ASR system.

3. SYSTEM ARCHITECTURE

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

All modules, except the audio/ASR module are implemented in the Java programming language. Both

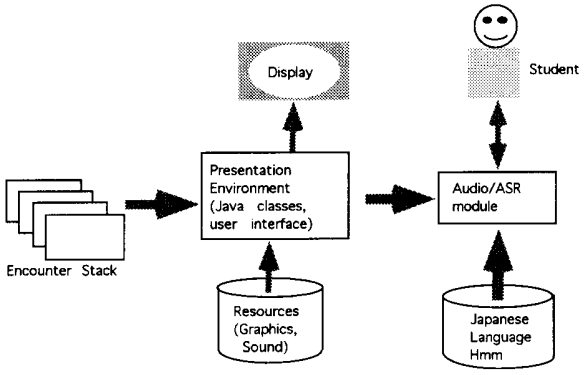

Figure 1: System Overview

the presentation environment and the audio/ASR module are 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 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. 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. The card object class, while serving the needs of an encounter-type exercise, is designed to be flexible enough to allow for any kind of spoken language exercise. The author of a particular encounter creates a recognition grammar for each occasion of verbal input from the student. This recognition 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 monophone 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 has been designed to be speaker independent and to accommodate a wide variety of non-native accents.

Although a beginning language student has limited proficiency, there is still a variety of potentially valid utterances that the student can produce in any situation, even if some of these may be prescriptively incorrect. Therefore, each encounter was prototyped in a traditional Hypercard environment on a Macintosh with text input and output only. Granted that text responses (right or wrong) will not be identical to students' spontaneous spoken responses, we assumed that they are probably similar. Hypercard provides an efficient means of modifying each encounter as the result of actual input from a test group of students.

4. TASKS

The current implementation of Subarashii has four encounters, listed here in the order of complexity and difficulty.

- Glad to Meet You (GTM)
- Movie Friday? (MF)
- Are you Busy? (AYB)
- Got Milk? (GM)

After selecting an encounter, a starting screen appears. It has the opening graphic display and a written mission statement. As students experience the encounters, they progress from a very passive encounter "Glad to Meet you" where they only respond to system-initiated prompts, to "Got Milk?" where the student has to take all the initiative.

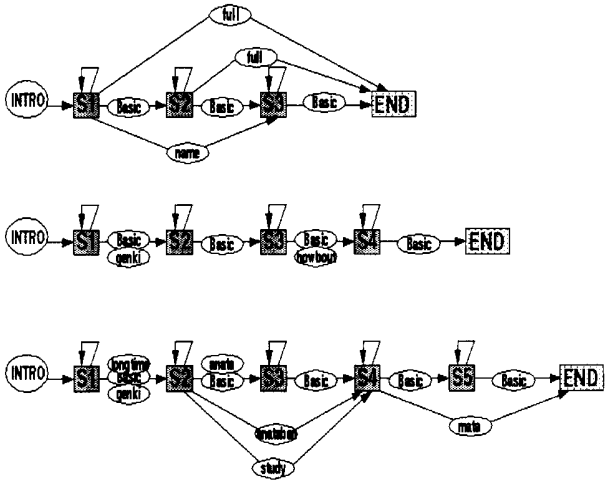

Figure 2: Network Schemata, First Three Encounters

Figure 2 shows the state diagram of the first three encounters; the fourth is much more complicated, even at the schematic level shown in Figure 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 elicits a different response from the system, but all of them lead 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 Figure 3. In this figure, ovals represent Japanese spoken by the system, rectangular boxes represent possible Japanese utterances by the student, and bold rectangular boxes represent written English comments to the student.

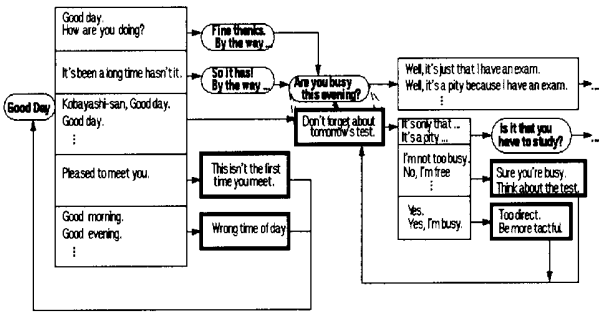

Figure 3: Initial Network for "Are you Busy"

5. EVALUATION

The Subarashii system was set up in 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. The sample of students mostly had A or B grade averages, along with 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 question 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.

5.1 Encounters

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 which were not rejected by the system. These were silent responses or problems with starting the recognizer which result in cut-off responses. 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 breakdown for the 38 sessions. Specifically, it shows the percentage of *in-network*, *out-of-network* and *operator-error* phrases. The latter two are further subdivided to *correct grammar usage* or *incorrect grammar usage*. Finally, the percentage accuracies of the system for *in-network* phrases are shown in parentheses.

On the average, only 2.6% of the time were the grammatically incorrect *in-network* phrases used 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 54% 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 doesn't seem to be a decrease in recognition accuracy with the more difficult encounters. In fact, we didn't see any significant correlation between grade average, number of years studying Japanese, or general interest and either recognition accuracy or correct grammar usage. We may assume, therefore, that even the fourth encounter is within the competence of first year students of Japanese.

Table 1: User and Recognition Behavior

	In-Network		Out-of-Net		Error
Grm:	Correct	Incor	Cor	Incor	All
GTM	54.2(94.6)	0(0)	1.5	39.1	5.2
MF	47.8(77.1)	2.4(100)	8.8	35.0	6.0
AYB	65.4(88.6)	7.3(33.3)	13.8	10.2	3.3
GM	55.0(77.3)	0(0)	29.0	12.9	3.0
Total	54.6(80.8)	2.6(55.6)	13.3	24.9	4.6

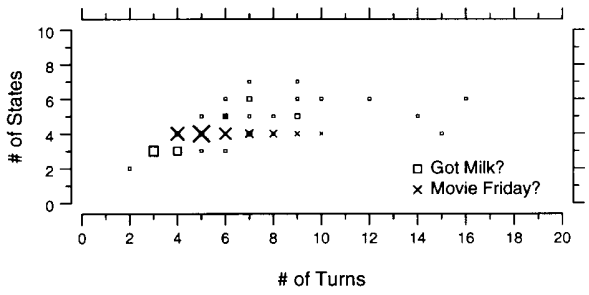

Figure 4: States Traversed vs. Turns Taken

We also looked the number of dialog turns (utterances) each student took to complete each task. Figure 4 shows a plot (for the *Got Milk?* and the *Movie Friday* encounters) of the number of turns that each student took to complete the encounter versus the number of machine-response states traversed (e.g. S1, S2, etc. in Figure 2). The size of the symbols indicates the number of students. Most of session 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.

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 the average, a student takes 1.4 turns to go through each state. This means that about 70 out of every 100 utterances advance the student to the next state. These advances consist of correctly recognized correct phrases and incorrectly recognized (as correct)

Table 2: Number of Turns and States

	Num of Turns			Num of States		
	Med	EV	SD	Med	EV	SD
GTM	4	4.6	2.3	3	3.0	0.6
AYB	5	5.1	1.4	5	4.7	1.0
MF	6	6.6	4.7	4	4.0	0.7
GM	6	6.6	3.6	4	4.4	1.5
Total	22	22.8	6.9	16	16.1	3.0

incorrect phrases. Table 1 shows that 54.6% of the utterances have both correct grammar usage and are in the grammar network, although only 81% of them are correctly recognized by the system. This means that only 44% the phrases are correctly recognized correct in-grammar sentences. If 70% of the utterances cause an advance to the next dialog state, but only 44% follow designed paths, 26% of the utterances that move the dialog forward still need to be accounted for.

Thus, the functional behavior of the system is better than its design warrents. 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, and we get a radically different view as shown in Table 3.

Table 3: Functional Behavior

	In Net		Out of Net		Combined
Grm:	Cor	Incor	Cor	Incor	Total(Incr)
GTM	95.9	0	50.0	62.0	72.8(+23.2)
MF	84.1	100	87.5	65.3	71.0(+33.5)
AYB	95.5	56.1	89.5	78.6	85.8(+26.4)
GM	84.8	0	54.2	41.3	67.8(+27.6)
Total	87.1	68.9	65.5	61.3	69.7(+25.7)

Table 3 shows the functional accuracy in each category as well the combined total. The number in parentheses in the last column indicates the serendipitous increase in functional accuracy as compared with the previous accuracy measurement. In this case, the percentage of correctly recognized correct sentences is increased to 56%, which indicates a more reasonable 14% false acceptance rate.

5.2 User Survey

We conducted two user surveys, one before and one after the students' experience with Subarashii. Neither revealed any significant correlation between survey results and performance with the encounters. The post-Subarashii survey solicited agreement or disagreement (with a response range from 5 to 1, respectively) with various assertions.

Students indicated high levels of comfort interacting in Japanese (average 4.1) and expressed confidence that they understand the computer (4.4), but were less confident that the computer understood them (3.9). Students generally agreed with the statement that they are better speakers of Japanese as a result of using the system (3.6). They wanted to use the program again (4.7) and thought that it could further improve their Japanese if used again (4.5).

6. DISCUSSION

The current Subarashii system provides preliminary evidence that meaningful conversational practice can be authored and implemented and that high school student do find these encounters useful. On-line access to dictionaries or other resources might improve the system for some users.

Our original goal was to make a system where interactions work without a mouse or keyboard, that operate at an appropriate conversational pace, and that incorporate verbal strategies for resolving misunderstandings. Although slowed down by recognition errors and out-of-network phrases, the Subarashii system operated fairly quickly and seemed to gain general user acceptance.

This system will be further evaluated at in a trial with university students after the grammar networks are extended with reference to the Silver Creek High School data. Finally, we plan to improve the raw recognition accuracy and move the system to a more affordable hardware platform.

ACKNOWLEDMENTS

The contents of this paper were developed under a grant from the U.S. Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the U.S. Federal Government. The authors wish to thank Kathleen Egan and Michelle Wee of the Federal Language Training Laboratory for suppport in design and content, Chubu University, Japan, for providing audio and graphical material, and Jaime Hannefield and Rushton Hurley for their help in data collection and analysis.

REFERENCES

[1] Bernstein, J., Cohen, M., Murveit, H., & Weintraub, M. (1990): "Automatic Evaluation and Training in English Pronunciation" *ICSLP-90*, Kobe, Japan. pp. 1185-1188.

[2] Neumeyer, L., Franco, H., Weintraub, M., & Price, P. (1996): "Automatic text-independent pronunciation scoring of foreign language student speech," *ICSLP- 96, Vol. 3*, pp. 1457-1460.