



Figure 2: Simultaneous Understanding System Timing Diagram

speech recognizer's output is sent to the "Understanding" part which translates the user utterances into commands. These commands, which contain the user's request/information, are executed on the "User Frame" (containing already inputted information) by the "Perform Action" part of the system. Information can be added, deleted, or completely reset via various "Perform Action" commands. These commands are divided into the following groupings:

- WANT** A new piece of information is added (e.g. *I am leaving at 11:00*).
- CHANGE** An old piece of information is replaced by a new one (e.g. *I am leaving at 7:00 not 11:00*).
- DELETE** An old piece of information is deleted (e.g. *I am not leaving at 7:00PM*).
- UNDO** The most recently entered input is deleted (e.g. *No, that's not correct, Wrong!*).
- CLEAR** Information in a particular field is deleted (e.g. *Destination is wrong*).
- RESET** The "User Frame" is completely cleared (e.g. *Clear All*).

This grouping was done both to give user a large amount of flexibility as well as to help us figure out what kinds of speech patterns are absolutely required for a realtime recognition system.

The "Database Access" section uses the information in the "User Frame" to lookup the database. The results from this access are passed to the "System Response" section to generate the appropriate system output. Finally, by using the Display and speech synthesis, the user is made aware of system's understanding of his/her own inquiry and its appropriate response.

3. SIMULTANEOUS UNDERSTANDING

Two different speech dialogue systems have been developed. One of them is the base system already described in Section 1. In this system, the computer and the user take turns waiting for each other: the computer has to wait until the user statement is finished before it can start processing, and then the user has to wait until the computer is finished before he/she can speak again.

The other system is based on the "Simultaneous Understanding" paradigm[9] in which the computer can accept a new utterance while still processing the previous one. Figure 2 shows a process timing diagram of the new system. As soon as there is a short pause (300ms) in the user input, the system sends the recognition result to the "Understanding" part. The "Understanding" part in turn sends the parsed user utterance to the "Perform Action" section. The "Perform Action" updates the "User Frame" and the Display with the new information. However, the database is not accessed yet, since even in medium-sized databases access time would take on the order of hundreds of millisecond, which is too long to wait. Finally, computer control is again returned to the "Understanding" part which awaits the next input.

This sequence of events continues until the users does not speak for a period of at least 1sec at which point the database is accessed, the display is updated with the new database information, and a system response is spoken via the speech synthesizer. Presently, while the speech synthesizer is working, it cannot accept and input speech from the user.

In this system, not only can the users speak several utterances in succession (as long there is a short pause between the utterances) stopping only to see the database results, but also if they see a mis-recognized entry, they can correct it before the database is accessed. However, how much difference these advantages make in the performance of the system is not very clear. Therefore, the following task was designed to measure our system quantitatively.

4. TASK

A Shinkansen (high speed bullet train) ticket reservation task was implemented to test our dialogue systems. In order to clearly show all the important information, the screen was divided into 3 parts. The top part shows a history of the user input and the system responses. The middle part displays information about the current contents of the "User Frame." Finally, the bottom portion shows the various train route information and ticket availability. The authors have also created a Shinkansen database that contains arrival and departure information about every Hikari and Nozomi train (Hikari & Nozomi are two kinds of Shinkansen trains). The system also contains information about the kind of seat availability (smoking/non smoking) and the position of the

seat (Reserved/not reserved/Green/cabin). A random number generator was used, so that on the average 90% of the seats are full as the create a more realistic setup. The recognition set for this system is a vocabulary set of about 290 words with a word perplexity of 8.0. Figure 3 displays what might be a typical user/system session.

```

USER> Hello.
SYSTEM> Hello. Where would you like to go?
USER> Do you have any train after 10 o'clock that goes
      to Osaka, ahhh, sorry, to Shin-kobe.
SYSTEM> Would you like Hikari-109?
USER> I would like Hikari-78...in the First Class section.
SYSTEM> There is no Hikari-78 going from Tokyo to
      Shin-kobe.
USER> Then please give me Hikari-79.
SYSTEM> How many tickets would like?
USER> 4 tickets, in the non-smoking section.
SYSTEM> Would you like to reserve this?
USER> Yes.
SYSTEM> Thank you very much.

```

Figure 3: Typical System Session

Notice that the user can both respond to the system inquires and/or make his own inquiries. The questions that are asked by the system are designed only to help the user with the kind of information that is still needed by the system in order to purchase a ticket, and it does not effect the recognition results.

Finally, the system always tries to come up with one suggestion for a train that fulfills all of the user's limitation (or it will inform the user if there aren't any).

5. EXPERIMENT

In order to compare the two systems, the authors came up with a set of typical situations that the system could be used to solve. A set of 5 different scenarios were used for the testing. Two of these could easily be solved using this system (e.g. the customer was asked for tickets for traveling from one city to another leaving at a particular time); while two could not be easily solved by the system (e.g. the customer was asked for the best tickets less than a certain amount, however price information cannot be inputted directly). The final case was one of user's own choice.

The recognized user input and the resulting "User Frame" information were saved on log files. The user entries were also recorded on tape and later transcribed them on file.

Eleven test subjects were used in the experiment. For testing the base system, the subjects were grouped in the following manner: 3 were completely unfamiliar with the system, 6 had used the system a few times before, and 2 people had used the system a lot. The 2nd experiment set on the new system was conducted on almost the same group of people except that one of the new users was replaced by someone who had some experience with the system before. Since there was a 11 month gap between the two experiments, it was assumed that the new users would not remember enough of their previous experience to be biased.

Table 1: Base System Results

Usr Grp	Total Sent	Corr	Mean Corr	Not Corr	Unkn Word	User Mist	Rejc
None	169	30.2%	23.7%	32.5%	5.9%	4.7%	3.0%
Some	509	30.8%	15.5%	40.1%	6.5%	2.2%	4.9%
Lot	168	38.7%	13.7%	42.9%	1.8%	2.4%	0.6%
All	846	32.3%	16.8%	39.1%	5.4%	2.7%	3.7%

Table 2: Simulated Understanding System Results

Usr Grp	Total Sent	Corr	Mean Corr	Not Corr	Unkn Word	User Mist	Rejc
None	158	50.0%	13.3%	23.4%	0.6%	9.5%	3.2%
Some	704	40.6%	8.9%	36.5%	8.5%	4.4%	1.0%
Lot	169	54.4%	12.4%	32.5%	0.0%	0.6%	0.0%
All	1031	44.3%	10.2%	33.9%	4.6%	5.9%	1.2%

The users were given brief instructions describing the general purpose of the system, the kind of information they could request, and less than a dozen examples of some user inputs. They were also given a questionnaire asking about the various aspects of the system performance. The authors used exactly the same instruction sets, scenarios, and questions for the two sets of experiments, without giving any directions to user about systems' differences.

Finally, the amount of time it took each subject to finish his/her scenarios was measured. However, the subjects were not told that they were being timed, or that they should try to go through the scenarios as quickly as possible.

6. RESULTS

In order to compare the two systems, the authors have found word recognition accuracy to be of little value since in some cases the authors observed that even though the meaning of the user utterance was correctly understood by the "Meaning Understanding" part, more than one or two words were mis-recognized by the "Speech Recognition" part. Therefore, the authors decided to base our accuracy measurements on phrase accuracy as opposed to word accuracy.

Hence, the input utterance have been categorized into one of six classifications: *Correct* for correctly recognized inputs, *Meaning Correct* for correctly understood input (but with one or more words of mis-recognition), *Unknown Words* for mis-recognized inputs with unknown words, *Not Correct* for mis-understood inputs (general), *User Mistake* for utterances in which the user spoke too early or spoke utterance not meant for recognition (e.g. user speaking to herself), and *Rejection* for those phrases that were rejected by the "Speech Recognition" part. Table 1 shows the results of the base system using the above classification. Table 3 shows the results of the "Simultaneous Understanding" system. In both cases, the results for all 11 users have been shown, as well as the separate results for the new user (*None*), the users with some experience (*Some*), and the user with a lot of experience (*Lot*).

As can be seen by comparing Table 1 & Table 2, the average accuracy (sum of *Correct* & *Meaning Correct*)

Table 3: Command Types Results

Cmd Type	WANT	CHNG	DELE	UNDO	CLEAR	RESET	UNCLS
Base	81.8%	0.2%	0.5%	5.5%	0.6%	8.8%	2.5%
Simul	80.8%	0.6%	1.1%	1.7%	0.9%	8.0%	4.9%

of the "Simulated Understanding" system was greater than the base system by 9.4% for the new users, 3.2% greater for the users with some experience, and 14.4% greater for users with a lot of experience. The total average for all 11 users, was 5.4% greater.

The authors looked at the the variations in usage of different command type in each system. As shown in Table 3, the results for 5 of command classifications are almost exactly identical. "WANT" type command was used about 81% of the time in each systems. Furthermore, "CHANGE," "DELETE," and "CLEAR" type commands seem to comprise only 1.3% of one system and 2.6% of the other one. However, there seems to be a large difference in the usage of "UNDO" commands and in usage of unclassifiable commands. The unclassifiable commands ("UNCLS" in Table 3) were almost all mistakes made by the user.

The authors also looked at the number of entries and the time spend by each user to solve the tasks. There were great variations among users, and no clear correlation could be seen to the amount of experience each user had. However, for the base system, it was found that each user spent an average of 785 seconds and spoke 77 entries to solve all 5 cases; or rather he spent 10.2 seconds per entry. For the "Simultaneous Understanding" system each user spent an average of 748 seconds and spoke 94 entries or 7.98 seconds per entry. This means an average of 21.8% reduction in the time spent per entry, and an average 4.7% reduction in the total time spent,

Furthermore, the authors also measured the average number of words in each user phrase. The average number was 8.8 words per utterance for the base system and 7.2 words per utterance for the "Simultaneous Understanding" system. This means that the average utterance has shortened by 18.2% between the two systems.

Finally, the responses to the questionnaire were compared. In both sets of responses, there were many comments about the the database information, the display, functions provided, and other aspects of the system. However, the most significant result was that for the base system 9 of the 11 users had some form of comment/complaint about the poor quality of recognition and the number of trials it takes to enter data; while in the "Simultaneous Understanding" system there were only 2 comments.

7. DISCUSSION

The authors have presented a new way to do interactive speaker recognition and have shown that it improves the average utterance accuracy by 5.4% and the average total time spent by 4.7%, These values, perhaps, may seem not to be significant enough to warrant a real difference between the systems; however, the authors believe that there are other issues involved in the analysis of the test results.

The authors have also studied the kind of inputs that are employed by the users and found very little difference in the usage of the commands types. However, it is noticed that the "UNDO" command was used less in the "Simultaneous Understanding System" which could be one of the reason for the increase in speed. However, at the same time, the authors note that the number of unclassifiable commands has increased.

One of the characteristics of our test experiments was that the scenarios were designed to test the user interactivity; the user is always given a number of trains and/or times to choose from. However, with this setup, it was hard to judge how well the users had done in solving the task. It was noticed that those users who had recognition problems tended to accept the computer's suggestion more quickly than those who had better recognition. Therefore, the latter group, although better at the recognition task, tended to repeat a mis-recognized entry more often and have a lower recognition score than it could have been had they always took the computer suggestion. Having a more restricted task would have alleviate this problem.

Although the results reported in this paper are preliminary, they show that the new method based on "Simultaneous Understanding" is promising for the "interactivity" of the speech dialogue system. Furthermore, the authors believe that a even a greater improvement can be shown if more restricted scenarios had been used.

ACKNOWLEDGMENTS

The authors wish to thank the members of the Human Language Research Laboratory for their continuous support.

REFERENCES

- [1] M. Bates, et al. "The BBN/HARC Spoken Language Understanding System." *ICASSP Proc* Apr. 1993, P. Vol, 111-114.
- [2] J. Polifroni, L. Hirschman, S. Seneff, V. Zue, "Experiments in Evaluating Interactive Spoken Language Systems," *Proc of Speech & Nat. Lang. Workshop* Feb. 1992, P. 28-33.
- [3] R. Pieraccini, et.al. "A Speech Understanding System Based on Statistical Representation of Semantics," *ICASSP Proc*, Mar. 1992, Vol I, P.193-197.
- [4] W. Ward, S. Issar, "Speech Understanding in Open Tasks," *Proc of Speech & Nat. Lang. Workshop* Feb. 1992, P. 78-83.
- [5] Y. Takebayashi, Y. Nagata, & H. Kanazawa, "Noisy Spontaneous Speech Understanding Using Noise Immunity Keyword Spotting," *ICASSP Proc* Apr. 1993, P. Vol II, 115-118.
- [6] K. Yoshida, T. Watanabe, & S. Koga, "Large Vocabulary Word Recognition Based on Demi-Syllable HMM," *ICASSP Proc.*, May 1989, P. 1-4.
- [7] K. Hatazaki, et al. "Intertalker: An Experimental Automatic Interpretation System Using Conceptual Representation," *ICSLP Proc.*, Oct. 1992, P. 393-396.
- [8] S.Koga, et al. "A Real-time Speaker-Independent Continuous Speech Recognition System Based on Demi-Syllable Units," *Proc. of ICSLP Proc*, Oct. 1992, P. 1483-1486.
- [9] K. Hatazaki, et.al, "Speech Dialogue Interface by Simultaneous Understanding," *Japanese Acoustical Society Proc.*, Mar. 1993, P. 75-76.