PROMPT CONSTRAINED NATURAL LANGUAGE - EVOLVING THE NEXT GENERATION OF TELEPHONY SERVICES

Stephen M. Marcus, Deborah W. Brown, Randy G. Goldberg,
Max S. Schoeffler, William R. Wetzel and Richard R. Rosinski

AT&T Business Communications Services, Holmdel, New Jersey, USA

ABSTRACT

This paper describes the design and development of an automated car reservation system using large vocabulary natural language speech recognition. Reservations were made over the public switched telephone network by calling an 800 number from anywhere in the United States, and car availability checked in real-time with a major international car rental company.

This system was designed to support first time users without requiring extensive written instructions, using current state-of-the art product-grade speech recognition technology. A Prompt Constrained Natural Language (PCNL) paradigm was used to encourage users to respond within the constraints of the recognition technology.

The results of an extended technology trial demonstrate the viability of services using automatic speech recognition to support transaction processing services. The extension of this technology to new domains will be described.

1. INTRODUCTION

The telephone network imposes many constraints on the use of automatic speech recognition. These include limited bandwidth, wide variations in transmission and handset characteristics.

The purpose of the "AutoRes" technology trial was to demonstrate that commercially available large vocabulary speech recognition systems have evolved to the point where they can be used to automate some of the tasks performed by human agents. This is particularly attractive for many national 800 telephone services, in which high and often unpredictable call volumes are concentrated on pools of agents.

Where appropriate computer-telephony integration is available to transfer collected information to an agent, even partial automation can result in valuable agent productivity enhancement. For example, if the system can determine a caller's account number and date of travel, but is unable to recognize their desired rental location, an agent could take over the interaction from that point onwards.

2. PROMPT CONSTRAINED NATURAL LANGUAGE

Simple keyword recognition systems are already automating simple tasks. Examples include operator automation, name dialing, and AT&T's "800 Speech Recognition" which allows callers to either "press or say" a digit for call routing.

The most basic of these systems use isolated word recognition - the user must speak only an appropriate keyword. Appropriate use of word-spotting technology, which models probable context phrases and/or contains "garbage" models for out-of-grammar utterances, substantially increase the usability of such systems.

Isolated Word Recognition	
Please say the month	*January*
What day of the month?	*15th*
Prompt Constrained Natural Language	
When will you be renting?	*January 15th*
	Tomorrow
	Monday, January 15th
Unconstrained Natural Language	
Wizard tours, how can I help you?	*I'd like a car for my trip tomorrow morning.*
	This is Stephen Marcus, KZ111778, I'd like a full size car at Logan Airport 10 a.m. tomorrow, arriving United flight 89.

Table 1: Illustrating Prompt Constrained Natural Language as a stepping stone from isolated word recognition to unconstrained natural language understanding.

Complex laboratory systems, such as the ARPA-funded Airline Travel Information System (ATIS) implementations, attempt to understand unrestricted natural utterances from a particular domain of discourse - in this case airline flight reservations. A typical utterance might be*: "I'd like the first United flight from San Francisco to New York tomorrow, returning on Saturday"* or, subsequently, *"What are the meals?"* These systems provide impressive performance in a laboratory setting, but are not yet accurate enough for use with real customers over the telephone network. Work is proceeding to (i) improve recognition accuracy

(ii) optimize systems for telephone-quality speech and (iii) design dialog and data selection strategies to replace the (often verbose) feedback currently given to the user using a graphic display with information that could be provided using speech over the telephone.

The current project is focused on developing a system using speech recognition technology available now to partially or fully automate a travel reservation service for an AT&T business customer. In order to provide the maximum flexibility to the user, we have adopted a middle road between isolated word recognition and unconstrained natural speech, one that we have termed Prompt Constrained Natural Language (PCNL). In this, specific prompts are designed to elicit responses from the user that current recognition technology is able to handle.

Table 1 gives an example of isolated word recognition, PCNL and unconstrained natural language determining essentially the same information from a speaker. This is, of course, only a static snapshot. As speech technology improves to allow accurate recognition of more complex utterances, PCNL prompts will be made more open to allow users the option of giving more information in response to a single prompt. Conversely, within a particular application, it may be found necessary to increase the constraints a prompt imposes on a user, even down to the point of requesting an isolated word response, in order to reliably solicit the desired information. We believe that it is this incremental approach, rather that directly attempting to solve natural language recognition, which is most likely to both provide solutions of short term value for AT&T products and services, *and* to drive technology the fastest towards less-constrained, and finally unconstrained, recognition. In addition, though extreme care should be taken in modeling automated systems after human dialogs, callers are often used to providing information in a sequential and constrained fashion, prompted by a human agent.

Note that the final example in **Table 1** is something that a human agent would not be able to handle, not because of limitations in human speech recognition, but because of limitations in human memory. It is therefore not unreasonable to suppose that improvements in automatic speech recognition and language understanding will one day result in systems which are able to perform at least such routine tasks more rapidly and accurately than a human listener.

In addition to careful design of each initial prompt for a particular item of information, successively more constraining reprompts were developed to encourage the user to respond appropriately after time-out or rejection by the recognizer. The use of less constraining (and consequently briefer) initial prompts avoids the problem demonstrated by Hone and Baber (1995) that although more constraining prompts may result in more appropriate user responses and less need for reprompting, they nevertheless increase total transaction time. In our design, initial prompts could be kept relatively short, though adequate for the majority of first-time users, while users who required more detailed instruction therefore received it without encumbering the dialog for other callers.

3. TRIAL DESCRIPTION

An iterative prototyping approach used to develop and refine the design of the system. Each involved a number of AT&T employees who had no prior involvement with the prototype calling in and attempting to make an automated car reservation.

The results of each trial were used to improve the system for the next trial, and ultimately for the online reservation system itself.

All trials were attempting to collect the following set of information:

- The caller's account code
- The caller's last name
- Originating location (250 major airport locations)
- Originating date
- Originating time
- Returning to same location? / (if no) Return Location
- Return Date

For all calls, the caller's spoken responses and detailed timing of all events were preserved in addition to the results of the speech recognizer. Responses were transcribed off-line. Recognition results were categorized as:

- **Correct**
- **Incorrect - in grammar** (misrecognition): an utterance in the recognizer's grammar was misrecognized as another utterance.
- **Incorrect - out of grammar** (false acceptance): an utterance outside of the recognizer's grammar was recognized as an in-vocabulary item.
- **Correct Rejection**: rejection of an out-of-grammar utterance
- **False Rejection**: rejection of an in-grammar utterance

In a task such as the current one, where multiple items of information need to be collected, incorrect recognition is much more serious than false rejection. In the former case, confirmation and repair dialogs are required to correct the error, whereas in the latter the system can simply reprompt for the desired information.

3.1 Trial A

The purpose of this trial was to collect data on users' spoken responses to an automated reservation system. A first version of the system was developed using only the team's intuitions about appropriate prompts and recognizer grammars. No feedback on recognizer results were given to the caller. Callers were solicited by email from within the AT&T research and business communities, and participants were asked to call in to an automated system and pretend they were reserving a car at a major airport location.

3.2 Trial B

The system was redesigned using the results from Trial A. Prompts were modified, the recognizer grammars were updated and the system was updated using new recognition software and a

substantially improved hardware configuration. The opportunity for confirmation and correction of the recognition results was also introduced into the dialog in this trial.

Participation was solicited in a similar manner to Trial A. The speech data from Trial B was both analyzed directly and used in a number of off-line experiments intended to improve recognition performance. These experiments resulted in the recognition models (grammars, rejection parameters and acoustic models) used in Trial C. Users' responses were also used to critically tune the prompts and develop appropriate reprompts.

3.3 Trial C

A number of significant enhancements to the Trial B system were made:

- Prompts and reprompts were redesigned
- The acoustic models used by the recognizer were augmented using a large AT&T subword data collection.
- Recognition grammars were modified based of the results of Trial B.
- New rejection models were used to substantially reduce the false-accept rate for out of grammar utterances.

Callers were again solicited from the AT&T community, but this time the system's telephone number was not circulated. Subjects from earlier trials were excluded. Participants were registered, assigned an account code, and sent four fictitious business meeting scenarios. They were asked to make flexible travel plans around these meetings, and to call in and make an automobile reservation for each trip separately. This allowed us to study how performance improved as a user gained experience with the system.

4. TRIAL RESULTS

4.1 Informal Observations

One surprising result of the trials was the extent to which users were willing to spontaneously adapt their language behavior to match their expectations of the system's capabilities. For example, our initial location grammar for Trial A included a number of phrases such as *"I will be renting at <location>"*. The most elaborate language behavior we observed in the actual trial were occasional ellipses (*"At Newark Airport"*) and ungrammatical hesitations and sentence fragments. The recognition grammars were therefore simplified to exclude complex phrases (and improve their performance and efficiency) and prompts which tended to evoke ellipses were reworded (e.g. asking *"What time?"* rather than *"At what time?"*).

Conversely, one prompt in Trial A was modeled after a query we observed used by human agents. Having determined the originating location, they often asked *"<location> pickup and return?"*, expecting in most cases the answer *"Yes."* In Trial A, we expected at best mediocre performance from our first cut at the recognition grammars, and thought such a prompt worth asking, expecting *"Yes"* or *"No"* when the location was recognized correctly, and some correction behavior when it was wrong. Our hope was to be able to use such implicit correction in a later version of the system. We found instead that the majority of responses were locations, dates and also times, with no simple way to even determine which locations were confirmations, corrections or alternative return locations. We therefore reworded this prompt in subsequent trials as *"Are you returning to the same location?"* and observed almost 100% simple affirmative/negative responses.

4.2 Recognition Performance

The recognition results for pickup and return locations are shown in **Figure 1** and illustrate the performance improvements as various components of the system were tuned during the trials.

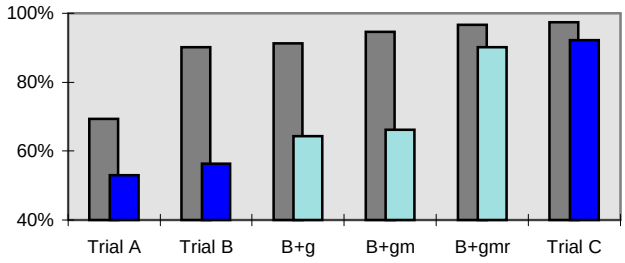

Figure 1 shows the improvement in percent correction recognition of locations in the various trials. The shaded bars represent recognition of in grammar utterances, the solid bars represent overall recognition. For a description of the off-line enhancements to Trial B (lighter bars), see text.

The various off-line manipulations which resulted in improvements in Trial B were as follows: (**B+g**) a new grammar was developed, based on the responses in Trial B itself. Not surprisingly, this resulted in better overall recognition, since more items are in the grammar. (**B+gm**) a new grammar, plus new acoustic models built using a large AT&T sub-word database. This resulted in almost halving the error rate on in-vocabulary utterances from 9% to 5%. (**B+gmr**) the further addition of a better rejection algorithm dramatically improved performance on out-of-grammar items. 87% of out-of-grammar utterances were correctly rejected, at the cost of a 9% false reject rate.

Trial C, which used all the enhancements from condition B+gmr with a new group of callers, confirms that the improvements resulting from the Trial B off-line modifications generalize to fresh data.

The results for location recognition are typical of the other data items collected, and show how careful application-specific design and tuning can take an initial "off the shelf" recognition score in the mid-50% range to over 95%.

4.3 Task Completion

Trial C contained all the major elements of the final reservation callflow, including confirmation, and, if necessary, correction of the collected information. The single most important measure of the total system is the percentage of calls in which all information is correctly captured. **Figure 2** shows the percentage of calls which were fully automated on the first, second and third calls by the same user. Reasonable first call performance of 50% full automation is followed by rapid improvement on the second and third calls. By the third call, the majority of calls which could not be fully automated had only one piece of missing information which would need to be completed by a human agent.

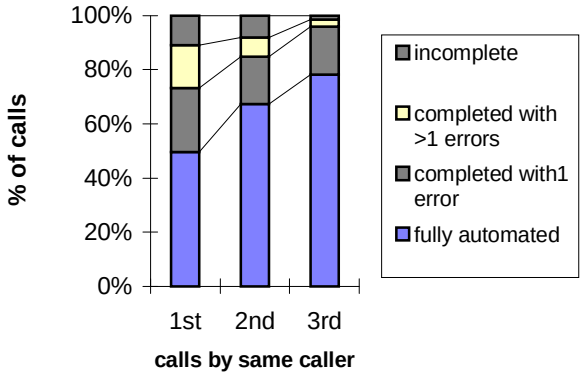

Figure 2 shows the improvement in call automation for successive calls by the same caller in Trial C.

5. FIELD TRIAL

Following the laboratory trials described above, a system essentially the same as the Trial C prototype was linked to the online reservation system of a major international car rental company. Acceptance testing was performed by 2,000 internal users making actual travel plans, followed by a field trial with 16,000 external customers.

The field trial lasted 6 weeks, during which time 375 calls were received. Most (73%) of these calls were from first-time users, with only a small number of calls from people who required two or more rentals during the period.

56% of the calls provided all the data required to attempt to reserve a car, a number very comparable to that found in Trial C. Of the remaining calls, a further 20% were requests outside the scope of the trial design (locations not included in the trial, users not amongst the trial participants, plus a small number of apparent test calls, where automation proceeded perfectly until the caller simply hung up). 20% of calls were routed to an agent because of one or more speech recognition errors, and only 3% of calls were hung up on because the user could not get a satisfactory response from the speech recognition system.

6. DISCUSSION

These results represent a snapshot of an ongoing development process. Recognition technology is continually improving, and recognition performance which was remarkable one or two years ago is now a commodity. At the same time, we are learning principles about human-computer dialog design which will both allow systems like the one described above to be improved, and guide the development of dialog modules and tools of general applicability to the automation of many routine transaction processing tasks.

7. CONCLUSIONS

The results of these trials show that a carefully designed, task-oriented human-computer dialog can be used to automate a significant number of calls which would normally be handled by a human agent.

The Prompt Constrained Natural Language paradigm has also been defined as a valuable tool for developing such large vocabulary speech recognition systems. Not only does it allow practical applications to be designed and built, but it also focuses research on the questions which will allow future extension of these systems to less constrained language domains.

8. REFERENCES

1. Hone, K.S. & Baber, C. “Using a simulation method to predict the transaction time effects of applying alternative levels of constraint to user utterances within speech interactive dialogues,” *ESCA Workshop on Spoken Dialogue Systems*: 8.1 209-212, 1995.