DESIGNING INTERACTIVE ERROR RECOVERY METHODS FOR SPEECH INTERFACES

Bernhard Suhm[1]*, Brad Myers*[2] *and Alex Waibel*[1]

[1] Interactive Systems Laboratories
Carnegie Mellon University (USA)
and
University of Karlsruhe (Germany)

[2] Human Computer Interaction Institute
Carnegie Mellon University (USA)

ABSTRACT

We present an approach to interactive recovery from speech recognition errors in the context of spoken language applications, which is motivated from research in the field of linguistics on repair dialogues in human to human conversations. We propose a framework to compare error recovery methods, arguing that a rational user will prefer error recovery methods which provide an optimal trade off between accuracy, speed and naturalness. We conjecture to beat skilled typing in speed, we need real-time speech recognition technology which performs on a 90% word accuracy level. Finally, we describe how we augmented our JANUS speech to speech translation system with prototypical error recovery methods, and presents results from a preliminary evaluation.

1. INTRODUCTION

Although intensive research over recent years has boosted the performance of speech recognition technology significantly, the automatic interpretation of speech is inherently unreliable - even human performance is limited [1]. We therefore argue that beyond improvements in baseline spoken language interpretation technology, methods to gracefully recover from interpretation errors are needed to be able to design speech user interfaces which live up to their promise of improving usability.

2. STRATEGIES TO DEAL WITH INTERPRETATION ERRORS

Research in the field of linguistics has investigated strategies people use in dealing with communication problems in conversations [2].

We argue this research can serve both as conceptual framework and source of inspiration for the design of error recovery strategies in spoken language enabled interfaces. The three main strategies employed in human to human conversation are avoiding communication problems, initiation of a repair dialogue as soon as a communication problem has been detected, and collaborative work on the repair [3]. Applied to spoken language interface design, these strategies correspond to the following three approaches to dealing with interpretation errors, some of which other researchers in the field have already addressed:

- Reduce the number of interpretation errors by influencing the user towards speaking styles the automatic interpretation system can interpret more accurately. [4],[5] showed that training the user can improve recognition accuracy. [6] explored how to use the interface layout and the system prompts to guide the user towards natural language queries which are easier to interpret.
- Facilitate the detection of interpretation errors through context sensitive feedback messages [7].
- Repair interpretation errors by involving the user in interactive error recovery.

Our research focuses on design of interactive error recovery methods.

3. DESIGN PRINCIPLES

We feel that the following dimensions are crucial in the design of spoken language interfaces in general, and error recovery methods in particular: application context, constraints of the current technology, and user preferences.

The application context varies along a set of dimensions:

- Learning time, ranging from virtually none in walk-up-and-use applications to elaborate training seminars.
- Bandwidth of interaction, ranging from uni-modal (e.g. speech-only telephone applications) to full multi-modal (e.g. Apple's vision of a Knowledge Navigator)
- Task, including dictation, (textual or visual) information retrieval and spoken language translation.

In human to human conversations, people prefer strategies which minimize the effort both communication partners spend collectively [8]. Therefore we argue the rational user will prefer spoken language interfaces which minimize the effort spent to complete his task, collaboratively with the spoken language interface. With respect to design of interactive error recovery, we conjecture a user will prefer methods which provide an optimal trade-off between

- Time required by the user to provide the input, and by the system to interpret it
- Interpretation accuracy
- "Naturalness" of interaction.

4. A PROTOTYPE INTERFACE WITH ERROR RECOVERY

We developed prototypical interactive methods to recover from speech recognition errors. Currently, we have implemented these methods in the context of a multi-lingual speech-to-speech translation system in an appointment scheduling domain [10]. Repair is performed at the level of speech recognition. The user interacts with the system until he is satisfied with the paraphrase of his input. Thus an interaction turn consists of providing an initial utterance, followed by - depending on the number of recognition errors - multiple repair interactions. The paraphrase is generated from the speech recognition output by translating the sentence back into the source language. Therefore, the goal of repair in this application context is to get a semantically equivalent sentence.

A repair interaction proceeds in two phases: First, an erroneous region in the speech recognizer's sentence hypothesis (the "reparandum") has to be identified. This reparandum identification can be initiated either by the system or by the user. Currently, we require the user to highlight the erroneous region. Then, the user can choose among various error recovery methods to correct the error:

- *Repair by respeak:* The user respeaks the reparandum, which is then replaced by the sentence hypothesis for the repair utterance. We developed language modeling techniques which take advantage of the (correct) context of the reparandum to improve repair accuracy.
- *Repair by spelling*: The user spells out loud a sequence of letters, which is recognized

by a specialized connected letter recognizer [11]. To achieve high accuracy, the words that are spelled can be constrained to some given vocabulary, e.g. the same as used by the continuous speech recognizer.

- *Repair by selection among alternatives*: A list of alternative hypotheses for the highlighted region is generated and displayed in a pop-up menu.
- *Repair by handwriting*: The user provides cursive handwriting input, which is interpreted by another specialized recognizer [12].

Currently, each repair method has a button associated with it, and the user initiates repair by pressing the corresponding button. This approach makes the choice of modality explicit and thus avoids moding errors. It remains to be explored whether eliminating all buttons by having the system determine automatically the appropriate specialized recognizer improves the repair interaction.

5. EXPERIMENTAL EVALUATION

Based on our most recent version of the JANUS speech translation system [10] we conducted a preliminary evaluation with 4 subjects. They had prior experience with speech recognition technology. In each session, two subjects were given fictitious calendars and were asked to schedule a meeting. Of a total of 57 turns, 39 needed repair. The initial decoding of a turn (with the continuous speech recognizer) yielded a word accuracy of 78%.

TABLE 1: shows the repair accuracies and on how many interations it was measured, for different error recovery methods. As can be seen, repair by spelling is the most accurate method, and selection among N-best alternatives the least. The latter reflects the fact that in most cases, no better alternative was found among the 10 best hypotheses. Additionally, the results show that our rescoring algorithm improves the accuracy of repair by paraphrase significantly.

	Respeak	-”- with Rescoring	Spelling	Handwriting	N-best Selection
Repair Accuracy	58%	66%	93%	85%	9%
# Interactions	29	29	15	20	37

TABLE 1: Repair Accuracies

Given the time-behavior (input and interpretation) and the accuracy of repair method, and a desired level of accuracy, we developed a method to estimate the speed to complete some input, including the time for repair. Based on the accuracy for the initial decoding and the repair accuracies above, TABLE 2: shows estimates for the speed to get 99% of the input correct - including time for repairs. To normalize for the length of the input we calculate the speeds in seconds/letter. The first two rows compare the speed using the currently most accurate ("3-pass") with a faster, but less accurate ("1-pass") version of our continuous speech recognition system. Surprisingly, although the 3-pass system outperforms the 1-pass system by 10% recognition accuracy (78% versus 68%), trading off speed against accuracy clearly speaks in favor of using the 1-pass decoder. Of course, the effect is most significant for repair by respeaking, which uses the continuous speech decoder not only to interpret the initial utterance, but also for repair utterances. Another surprising result is that the proposed additional rescoring pass overall slows repair down, although it increases accuracy and thus reduces the number of repair interactions. The estimate for N-best selection is somewhat bogus since the repair accuracy for this method is so low.

	Respeak	-"- with Rescoring	Spelling	Handwriting	N-best Selection
3-pass Decoder	6.6 [seconds/letter]	6.6	1.7	3.3	21.8
1-pass Decoder	3.8 [seconds/letter]	4.3	1.4	3.2	23.9

TABLE 2: Repair Speed

To explore the potential impact of improved interpretation technology on user preferences from a speed oriented point of view, we measured the average length of the input signal (also normalized by the number of letters) for the various modalities used, as shown in TABLE: 3. We consider this as a lower bound for the speed of different repair methods, i.e. assuming real-time and 100% accurate interpretation. Confirming intuition, continuous speech, i.e. repair by respeak, is the fastest modality, whereas handwriting is significantly slower.

Respeak	-"- with Rescoring	Spelling	Handwriting	N-best Selection
0.12 [seconds/letter]	0.12	0.5	1.3	0.6

TABLE 3: Lower Bounds on Repair Speed

These results suggest the following conclusions for the design of speech interfaces augmented with error recovery:

- With current technology, spelling is the most effective repair method.
- With further improving accuracy and close to realtime continuous speech recognition, we expect users to prefer repair by respeak.
- From the point of view of a rational user, it won't make sense to offer both speech and handwriting as alternative input modalities: Since handwriting is much slower, handwriting will be preferred only if it is the more "natural" input modality for the task at hand (e.g. filling out forms).

6. HOW FAR DO WE HAVE TO PUSH SPEECH RECOGNITION TECHNOLOGY?

Trying to envision the state of the art in spoken language technology in 2 years, one can expect close to realtime, but still unreliable interpretation of spoken language.

An interesting question is how far we have to push the accuracy of spoken language technology such that it will be competitive with traditional input modalities, e.g. typing. Based on our method to estimate the overall speed for some input modality including repair, we derived estimates how accurate a continuous speech recognizer (CSR) would have to be in order to be faster than skilled typing. TABLE 4: shows the derived lower bounds on word accuracy. We note these estimates are optimistic since they assume realtime recognition and ignore unavoidable overhead, e.g. time necessary to press buttons or to plan the next step.

Typing Speed	Minimum Word Accuracy
40 words/minute	70%
60	84%
100	90%

TABLE 4: Lower Bounds on Word Accuracy for CSR to beat Typing

7. ACKNOWLEDGEMENTS

This research was sponsored in part by the German Ministery for Science and Technology (BMFT) as a part of the VERBMOBIL project and by the U.S. Department of Defense

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Navy or the U.S. Government.

8. REFERENCES

[1] W.J. Ebel,and J. Picone: "Human Speech Recognition Performance on the 1994 CSR Spoke 10 Corpus", *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, 1995, pp. 53-59

[2] E.A. Schegloff, G. Jefferson and H. Sacks: "The preference for self--correction in the organization of repair in conversation", *Language*, 1977, Vol. 53, pp. 361-382

[3] H.H. Clark and E.F. Schaefer: "Contributing to Discourse", *Cognitive Science*, 1989, Vol. 13, pp. 259-294

[4] E. Zoltan-Ford: "How to get people to say and type what computers can understand", *International Journal of Man-Machine Studies*, Vol. 34, 1991, pp. 527-547

[5] Catalina M. Danis: "Developing Successful Speakers for and Automatic Speech Recognition System", *Proceedings of the Human Factors Scociety 33rd Annual Meeting*, 1989

[6] Sharon L. Oviatt, Philip R. Cohen and Michelle Wang: "Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity", *Speech Communication*, 1995, Vol. 15, pp. 283-300

[7] Susan E. Brennan and Eric A. Hulteen: "Interaction and feedback in a spoken language system: a theoretical framework", *Knowledge Based Systems*, 1995, Vol. 8, pp. 143-151

[8] H.H. Clark and D. Wilkes-Gibbs: "Referring as collaborative process", *Cognition*, 1986, Vol. 22, pp. 1-39

[9] Arthur E. McNair and Alex Waibel: "Improving Recognizer Acceptance through Robust, Natural Speech Repair", Proceedings of the International Conference on Spoken Language Processing - ICSLP, 1994, Yokohama (Japan), Vol. III, pp. 1299-1302

[10] B. Suhm, P. Geutner,T. Kemp, A. Lavie, L. Mayfield,A. McNair, I. Rogina, T. Schultz, T. Sloboda, W. Ward, M. Woszczyna, and A. Waibel: "JANUS: Towards Multilingual Spoken Language Translation," *Proceedings of the ARPA Spoken Language Systems Workshop,* 1995

[11] H. Hild and A. Waibel: "Speaker-Independent Connected Letter Recognition with a Multi-State Time Delay Neural Network", *3rd European Conference on Speech, Communication and Technology* - EUROSPEECH, Berlin, Germany, September 1993, pages 1481 - 1484

[12] S. Manke, M. Finke, and A. Waibel: "NPen++: A Writer Independent, Large Vocabulary On-Line Cursive Handwriting Recognition System", *Proceedings of the International Conference on Document Analysis and Recognition*, Montreal, 1995