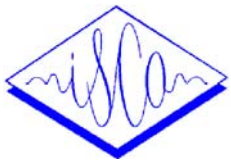


EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

EXPLOITING REPAIR CONTEXT IN INTERACTIVE ERROR RECOVERY

Bernhard Suhm and Alex Waibel

Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh PA, USA
University of Karlsruhe, Karlsruhe, Germany
Email: {bsuhm,ahw}@cs.cmu.edu

ABSTRACT

In current speech applications, facilities to correct recognition errors are limited to either choosing among alternative hypotheses (either by voice or by mouseclick) or respeaking. Information from the context a repair is ignored. We developed a method which improves the accuracy of correcting speech recognition errors interactively by taking into account the context of the repair interaction. The basic idea is to use the same language modeling information used in the initial decoding of continuous speech input for decoding (isolated word) repair input. The repair is not limited to speech, but the user can choose to switch modality, for instance spelling or handwriting a word. We implemented this idea by rescoring N-best lists obtained from decoding the repair input using language model scores for trigrams which include the corrected word. We evaluated the method on a set of repairs by respeaking, spelling and handwriting which we collected with our prototypical continuous speech dictation interface. The method can increase the accuracy of repair significantly, compared to recognizing the repair input as independent event.

1. INTRODUCTION

For any application of speech technology, the problem of recognition errors has to be addressed. In fact we believe the lack of graceful ways to recover from errors is a major reason that to date, speech recognition applications haven't quite met expectations. Commercial products are basically limited to isolated word recognition domains or small vocabularies, and success stories have been few.

Our approach to address the problem is to have the user interactively locate and correct errors, previously presented in [1],[2]. We assume a user willing to collaborate with the interface in correcting errors as long as he can complete his task that way more efficiently. However, it is crucial that the chances for successful repair are higher compared to the trivial "try again".

We argue there are two ways to increase the probability for successful correction: first, by switching to another input modality, for instance from speech to spelling or handwriting, thus providing a signal orthogonal to the original, misinterpreted one. Second, repair accuracy can be increased by correlating the input the user provides in his attempt to correct with the repair context. While in prior work [2], we have explored the benefits of switching modalities, this paper presents a method which attempts to increase accuracy of repair by correlating repair input with the context of repair. A very weak form of such correlation is to eliminate words which the user identified as erroneous from the recognition vocabulary used during decoding of repair input. This idea corresponds to Ainsworth et al.'s "repair by elimination" [3]. Whereas Ainsworth considered unimodal repair (by respeaking) only, it can be applied to cross modal repair in a straightforward manner.

The remainder of this paper is organized as follows. Section 2 briefly reviews our approach of multimodal interactive error recovery. In section 3, we describe our method to correlate repair input and repair context. Finally, section 4 presents results based on data collected with our protoypical dictation interface.

2. MULTIMODAL INTERACTIVE ERROR RECOVERY

Although intensive research has significantly increased performance of speech recognition systems on certain benchmark tasks commonly adopted in the speech recognition field (e.g. Wall Street Journal, Switchboard), it still degrades dramatically in speech recognition applications. Additionally it is widely believed that recognition performance will remain limited. Therefore, further advances in speech recognition technology will not eliminate the need to address the problem of errors in the design of speech recognition applications.

We argue that speech user interfaces are feasible despite limited performance of speech recognition systems if the potential for error is balanced by efficient ways to correct them. Interactive error correction proceeds in two steps: error identification and error correction. Errors can be identified either by the system, for instance by highlighting words with low confidence scores, or by the user, for instance by selecting misrecognized words with a pointing device. For error correction, the user can choose among different correction methods, potentially switching input modality, for instance from continuous speech to spelling or handwriting.

There are several motivations for the multimodal approach to error recovery: First, without switching modality, accuracy of recognizing repair input is lower than the baseline accuracy, since misrecognized words tend to be inherently more difficult to recognize. Also, users may attempt to help the recognizer by hyperarticulating, a strategy often employed in human-to-human communication, making the recognition task even more difficult. Second, we can exploit that different input

modalities are orthogonal: words which are confusable in one modality can be disambiguated based on input in a different modality. Finally, recent studies by Oviatt [4] suggest that user frustration is alleviated by switching modality alone, regardless whether the chances for successful repair is higher in the new modality.

Multimodal interactive error recovery is adequate for speech applications which allow some form of graphical user interface, including a writeable display (e.g. touchscreen). In addition, the task should require the input to be recognized verbatim, so that it is natural for the user to focus his attention to a string of words (the displayed recognition hypothesis), for example in dictation applications. The situation is different for spoken dialogue applications, where the meaning of some (voice) input is sufficient to initiate some action which will satisfy the user's request. Also, the multimodal approach can be limited by hardware constraints of the application. For example, telephone applications typically require a speech-only interface, at the most enhanced with a very small display. There are initial attempts to address the problem of repair in spoken dialogue systems, for instance by Danieli et al. [5].

3. EXPLOITING REPAIR CONTEXT

Instead of interpreting repair input as an independent event, we propose to correlate it with the context of the repair. A very simple such correlation is to eliminate in repair attempts rejected words from the vocabulary used to decode subsequent repair input. Of course, after a repair has been completed, the original vocabulary has to be reestablished.

We developed a more powerful method to exploit contextual knowledge typical for the repair situation. It is based on the observation that the words in the vicinity of an identified error are correct. Note that albeit speech recognition errors typically do not occur isolated but in islands of consecutive misrecognitions, it is reasonable to assume that the user will correct errors starting from the beginning of a sentence, and will try to correct consecutive erroneous words in one repair if possible.

In the following, we will refer to the as misrecognized identified word(s) as *reparandum*, and to the words in the vicinity of the reparandum as *repair context*. To correlate repair input (provided to replace the reparandum) with the repair context, we use the following simple rescoring method: The repair input is decoded as an independent event, and a list of alternative interpretations is obtained. Then, we compute the language model score for each N-gram which contain at least one word from both repair context and reparandum. The list of hypotheses for the repair input is then rescored using an interpolation of these "context scores" with the recognition score, and the reparandum is replaced by the best choice from the rescored list.

For example, in our current prototype, we use a standard trigram language model, and repair input in modalities other than speech (i.e. spelling and handwriting) is limited to isolated words. In this situation, it is sufficient to consider the two words preceding and following some identified error. If we denote these words as w_{-2}, w_{-1} and w_{+1}, w_{+2} respectively, the context score *CSc(k)* for the k-best repair input hypothesis r^k is given by

$$CSc(k) = P\left(r^k \middle| w_{-2}w_{-1}\right) + P\left(w_1 \middle| w_{-1}r^k\right) + P\left(w_2 \middle| r^k w_1\right)$$

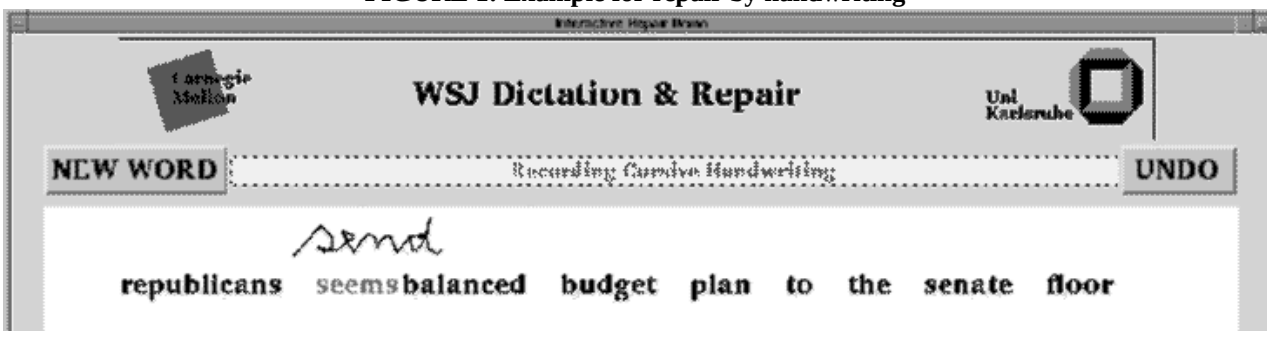
and the final score for r^k can be calculated as linear interpolation of context score with recognition score. This method can be extended in a straightforward manner to multiple word repair input, to statistical language models other than the standard N-gram, and to rescoring of lattices instead of (M-best) lists of hypotheses.

4. EVALUATION

4.1. Mulitmodal Error Recovery Interface for Dictation

We have implemented a prototypical speech user interface with multimodal interactive error recovery capabilities for continuous speech dictation. The user can dictate sentences using continuous speech and the recognition hypothesis is displayed. Currently, error identification is done by the user. In case of substitution errors, he simply highlights regions of misrecognized words in the displayed recognition hypothesis using a pointing device. In case of deletion errors, he can position the input cursor appropriately using intuitive hand-drawn pen gestures, which are motivated by gestures used in editing tasks with paper and pencil [7]. For error correction, there are different ways to address each of the three different kinds of recognition errors - insertion, deletion or substitution error. Inserted words can be deleted using a different set of hand-drawn pen gestures similar to the ones identified by Wolf et al. [7]. Substitution and deletion errors can be corrected by replacing the highlighted error or by inserting at the current position of the cursor, respectively. Currently supported repair input modalities are respeaking, spelling and handwriting. In addition, the standard correction method of choosing among alternative words is available. Figure 1 shows an example for repair by handwriting: the user spoke "republicans send a balanced budget plan to the senate floor", and corrected the misrecognized "send" by writing on the touchscreen.

FIGURE 1: Example for repair by handwriting



We process the different input modalities using specialized recognizers. The speech recognition subsystem is based on the JANUS recognition engine [8] in the configuration for large vocabulary Wall Street Journal dictation. Spelling input is processed by a specialized high-performance, real-time continuous spelling recognizer [9]. The pen input subsystem consists of a MS-TDNN-based handwriting recognizer capable of processing writer-independent, cursive handwriting of isolated words [10], a template-matching based gesture recognizer [11] and a simple heuristics to decide when to invoke handwriting versus gesture recognition on pen input. For all recognizers (except for the gesture recognizer) we use a standard 20K Wall Street Journal vocabulary.

4.2. Data

Using the above described interface, we collected multimodal repair interactions. Subjects had to dictate a given text from the Wall Street Journal in continuous speech, and then repair recognition errors using a modality of their choice. From interactions of 5 subjects we identified 42 instances of repair by (continuous) speech, 115 instances of repair by handwriting, and 97 instances of repair by spelling.

4.3. Results and Discussion

Based on the above data, we performed rescoring experiments, comparing three conditions:

1. treating the repair input as independent event (i.e. no rescoring)
2. rescoring using the (trigram) repair context preceding the error ("pre context")
3. rescoring using the trigram context both preceding and following the error ("pre and post context")

Table 1 shows the repair accuracies for these different conditions. As can be seen, correlating repair input with repair context could significantly increase accuracy for repair by speech and handwriting, where there was no effect for the spelling modality. The reason is that in the few instances where repair by spelling was misrecognized as independent event, the correct word was not in the N-best list of hypotheses.

TABLE 1: Repair Accuracies w/o Context Rescoring

	speech	handwriting	spelling
# repairs	42	115	97
1. independent event	35.7%	68.7%	92.8%
2. pre context	54.8%	80.9%	92.8%
3. pre+post context	52.4%	82.6%	92.8%

We hope to remedy this problem by extending the list of alternative hypotheses from the recognizer by additional words which can be considered "confusable" with those the recognizer found.

Since we were forced to trade-off speed against accuracy for continuous speech recognition to allow real-time interactive user tests, the performance of the baseline speech recognizer used was clearly suboptimal, performing at below 70% on test data of the official November'94 WSJ Hub evaluation. Therefore, the context following the error was frequently not correct, and using that context in addition to the context preceding the error did not yield consistent results. However, with increasing accuracy of the baseline continuous speech recognizer, we expect that rescoring with the context both preceding and following the error will outperform the pre context only method consistently.

5. CONCLUSION

The lack of graceful and effective ways to recover from recognition errors is one of the major links missing to make speech user interfaces more successful. Our multimodal interactive approach to error correction is promising for speech applications which allow for a graphical user interface, for example dictation. Exploiting information from the context of repair interactions is necessary to maximize effectiveness of repair. We have shown that using language model constraints from the context of misrecognized words to decode isolated word repair input can significantly increase repair accuracy.

6. ACKNOWLEDGEMENTS

This research was sponsored by the DARPA under the Department of the Navy, Office of Naval Research under grant number N00014-93-1-0806. 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.

7. REFERENCES

[1] McNair, A.E. and Waibel, A. "Improving Recognizer Acceptance through Robust, Natural Speech Repair", *Proc. ICSLP'94*, pp. 1299-1302, Yokohama, 1994

[2] Suhm, B., Myers, B. and Waibel, A. "Interactive Recovery from Speech Recognition Errors in Speech User Interfaces", *Proc. ICSLP'96*, pp. 861-864, Philadelphia, 1996

[3] Ainsworth, W.A., and Pratt, S.R. "Feedback strategies for error correction in speech recognition systems", *Int. Journal of Man-Machine Studies*, Vol. 36, pp. 833-842, 1992

[4] Oviatt, S.L. and VanGent, R. "Error Resolution during Multimodal Human-Computer Interaction", Proc. ICSLP'96, Philadelphia, 1996

[5] Danieli, M. "On the use of expectations for Detecting and Repairing Human-Machine Miscommunication", AAAI Workshop on Detecting, Repairing and Preventing Human-Machine Miscommunication, Portland, 1996

[6] Wolf, C.G. and Morrel-Samuels, P. "The use of hand-drawn gestures for text editing", *International Journal for Man-Machine Studies*, Vol. 27, pp.91-102, 1987

[7] Waibel, A. et al. "JANUS-II - Advances in Spontaneous Speech Recognition", *Proc. ICASSP'96*, Atlanta, 1996

[8] Hild, H., and Waibel, A. "Speaker-Independent Connected Letter Recognition with a Multi-State Time Delay Neural Network", *Proc. EUROSPEECH'93*, pp. 1481-1484, Berlin, 1993

[9] Manke, S., Finke, M., and Waibel, A. "NPen++: A Writer Independent, Large Vocabulary On-Line Cursive Hand writing Recognition System", *Proc. Int. Conf. on Document Analysis and Recognition*, Montreal, 1995

[10] Rubine, D. "The Automatic Recognition of Gestures", Ph.D. Thesis, Carnegie Mellon University, 1993