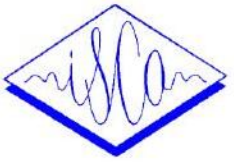

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

IMPROVING RECOGNIZER ACCEPTANCE THROUGH ROBUST, NATURAL SPEECH REPAIR

Arthur E. McNair and Alex Waibel

Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

10.21437/ICSLP.1994-339

ABSTRACT

Though large vocabulary speech recognition systems have improved greatly in recent years, usability of these systems in practical applications is still low, due to the ever-present errors. Adding a natural interface to these systems for users to correct errors should increase acceptance. This paper describes three methods for accomplishing speech-based repair of a misrecognition. The user must respeak or spell only an errorful subsection of the original utterance. A method is described to automatically locate the respoken subpiece in up to 90% of the instances. Once the subpiece has been respoken and located, another method is described which corrects the subpiece in up to 70% of the instances. If the location is known, and the subpiece is spelled, a third method is described which uses a single spelling utterance to correct the subpiece in up to 82% of the instances. The results indicate that these methods can decrease the error rate of a CSR by two thirds using only a single short repair utterance.

1. INTRODUCTION

Most current research in speech recognition focuses on the continuing improvement of large-vocabulary speech recognition accuracy. While great improvements have been made in recent years, no recognition algorithms or systems have been created which eliminate the possibility of recognition errors. If large vocabulary speech recognizers are going to be used for any tasks where exact recognitions are critical, then the inevitable errors need to be eliminated in some way that is acceptable to the users of the system. This means that a user interface must be designed to allow the user to correct recognition errors.

The simplest error correction interface is to force the user to respeak the whole utterance numerous times until the recognizer gets it correct. This interface may be easy to design and build, but it meets with very low user acceptance, due to the fact that greater user investment of time does not lead to a greater likelihood of the error being corrected.

Another interface design is to force the user to edit the recognized text with a keyboard or mouse-based editor. Though this method may guarantee correction of the errors, it requires the user to switch input modalities to accomplish a single task, and also eliminates many of the hands-free, eyes-free benefits of a speech interface.

A better recognition repair interface is one which allows the user to repair misrecognitions by voice only, in a way that is natural and effective in human to human communication. One of the most common human to human repair methods is to respeak a misrecognized portion of an utterance, speaking more clearly, and often hyper-articulating the word or word sequence which was misrecognized. If some word or words are particularly confusing, humans will often use spelling as a method of disambiguation.

In this paper, we describe methods used to implement a speech interface for repairing misrecognitions by simply respeaking or spelling a misrecognized section of an utterance. While much speech "repair" work has focused on repairs within a single spontaneous utterance [1], we are concerned with the repair of errorful recognizer hypotheses.

2. METHODS

To perform the techniques described in this paper, a continuous speech recognizer is required which can switch language models quickly. The recognizer must also be able to output a segmented, scored n-best list and/or word lattice. For this work, we used the LVQ-II continuous speech recognizer [2]. For spelling, we used an MSTDNN-based continuous spelling recognizer [3]. Both recognizers use the same input format, which simplifies their use in one common inter-

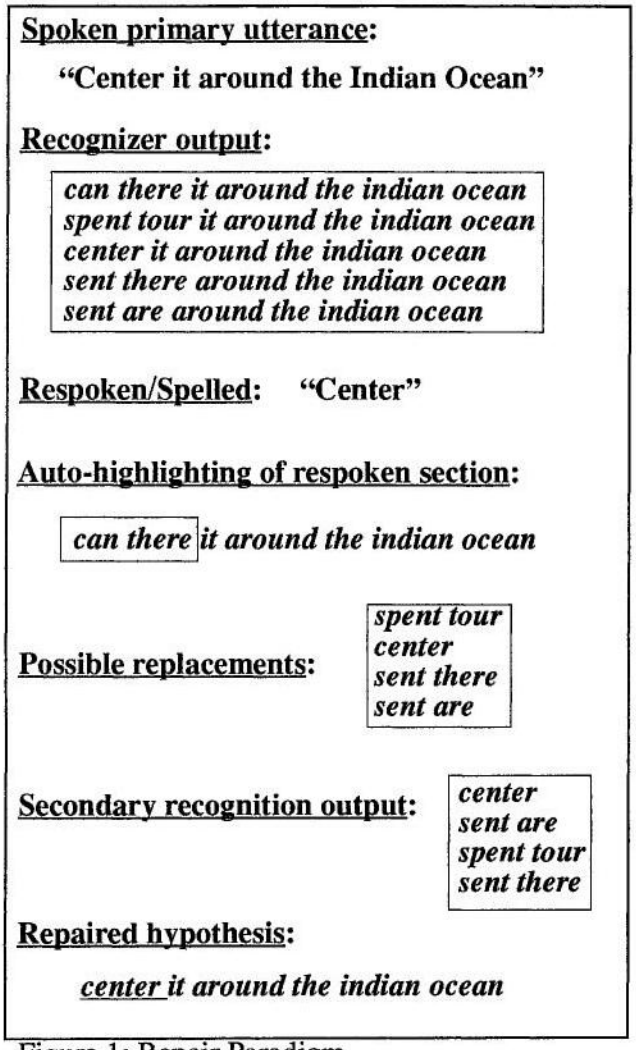

Figure 1: Repair Paradigm

face. The CSR system uses a bigram or wordpair language model to constrain the hypothesis search, while the spelling recognizer can use letter n-grams or finite state grammars.

Three methods will be described here: One, the automatic subpiece location method; Two, the spoken hypothesis repair method; and Three, the spelling hypothesis repair method.

Figure 1 shows the standard repair paradigm used. The speaker first utters a primary utterance, which is recognized in the primary recognition. If an error occurs, the speaker respeaks or spells the erroneous subsection of the primary utterance. This secondary utterance (or repair utterance) is recognized in the secondary recognition, using a language model constructed separately for the specific repair situation. The results of both steps are then used to locate and/or repair the original error.

2.1 Automatic Subpiece Location

This technique is used when a primary utterance has been spoken and recognized with an error. One of the necessary pieces of information to repair an error is the location of that error in the primary utterance. This location could be determined with the use of a mouse, highlighting an errorful subsection of the recognition. Here, we describe how to accomplish this highlighting by voice only, requiring the user to respeak only the errorful subsection of the primary utterance.

Given that the user will respeak some unknown subsection of the primary utterance, a language model is created which will allow all substrings of the first hypothesis of the primary recognition. The secondary utterance (a respeaking of a subpiece of the primary utterance) is then run through the recognizer, using this newly constructed language model. This will produce the secondary n-best list of possible choices for the respoken subpiece. Each hypothesis in the secondary n-best list (from best to worst) is evaluated to determine if it is a substring of the first hypothesis of the primary recognition. If it is a substring, the evaluation stops, and the endpoints of this substring are returned as the location of the respoken subpiece.

There is some possibility that no subpiece is found, since wordpair language models are not strong enough to guarantee this. A finite state grammar to constrain the search would be able to guarantee that only exact substrings are produced. In this set of experiments, wordpair models were found to be sufficient, always producing some subpiece within the first five secondary recognition hypotheses.

There is also the problem that there might be multiple, identical subpieces in the primary recognition first hypothesis. In this case, recognizing exactly what sequence of words was respoken is not enough to determine which of any identical sequences in the utterance was respoken. This problem would be most prevalent in commonly repetitive strings of numbers or letters. For the current experiments with the Resource Management task, the first matching subpiece (scanned in normal reading order) in the primary recognition hypothesis was used. Though other selection criteria could be used, this simple method was found to work well for this mostly non-repetitive RM task.

Preliminary testing of this method showed that it works poorly if the subpiece to be located is only one or two short words, as might be expected.

Though a great disadvantage in specific situations, this drawback is not seen much in actual usage, since humans tend to respeak a few words around the error to make it easier for other humans to locate the exact position in the utterance where the misrecognition occurred.

2.2 Spoken Hypothesis Correction Method

Assuming that the location of an errorful section is already known (highlighted), and the user has respoken it, how do we combine the information from the primary recognition and this secondary utterance to repair the original hypothesis?

The idea is relatively simple, but requires the assumption that the correct subpiece is in the n-best list (or word lattice) somewhere. A language model is created which restricts each hypothesis of the secondary recognition to be one of the alternative substrings in the primary n-best list at the same location as the highlighted errorful substring. The secondary utterance is then recognized using this new language model. In the simplest form, the top hypothesis from this secondary recognition can be used to replace the errorful subsection of the first hypothesis of the primary recognition.

In this experiment, the language model used was a simple bigram model (no unseen wordpair probability) based only on the counts found in the appropriate subpieces of the n-best list.

To find all the possible subpieces in the n-best list which were alternatives for the highlighted section of the best hypothesis, the start and end frames of the highlighted section were determined. In all other n-best hypotheses, the subpiece was chosen to include any words between or overlapping these start and end frames. Only unique substrings were used to determine the counts for the bigram language model. The original subpiece (known to contain at least one error) is also excluded from the language model data so that it cannot reoccur.

Merely replacing the errorful sub-section with the top hypothesis from the secondary recognition means that all of the subpiece order information from the primary n-best list is unused. To make use of this information, we use a method which rescores and reorders the secondary recognition list by averaging the scores from the secondary recognition list with scores of identical subpieces in the primary recognition list.

2.3 Spelling Hypothesis Correction Method

Repairing a recognition with spelling is very similar to correction by speech. Again, given the location of an errorful section, a secondary utterance and search can be used to repair the primary recognition. In this case, the secondary utterance is a spelling of the correct words for the subsection that contained the error. In this experiment, all the words are spelled together, with no break for word boundaries. A string of words like "GET ME ANY" would be spoken as "G-E-T-M-E-A-N-Y". Again, a language model is created from the subpiece hypotheses in the same position as the errorful subsection in the primary n-best list. For spelling, this language model is a finite state grammar, which completely restricts the output of the recognition to exact alternatives to the highlighted text. (The spelling recognizer does not produce an n-best list or letter lattice). The subpiece that was recognized by the speller is then used to replace the errorful subpiece in the original hypothesis.

Another method tried is to let the spelling recognizer do a free recognition (no language model), and then score each possible sub-piece by its dtw distance from the recognized sequence. This gives a score for each subpiece, which allows the combination of scores from the spelling recognition and the primary n-best list to come up with the best replacement subpiece.

3. TEST DATA

The ARPA Resource Management task was used for our experiments. The specific set of utterances chosen consisted of all the male utterances from the February and October 1989 ARPA test data. This included 390 utterances, in which were 300 unique sentences.

In experiment 1, the original ARPA speakers' utterances were used as the primary utterance, and, in those cases where recognition errors occurred, a separate speaker recorded both the respoken and spelled repair utterances.

In experiment 2, the same speaker spoke all 390 primary utterances as well as the respoken repair utterances for those primary utterances that were misrecognized.

For these experiments, the CSR recognizer was run in a sub-optimal mode, in order to generate more errorful tokens over our test database.

4. RESULTS

Table 1 shows the primary recognition accuracies for the CSR in both experiments. Table 1 also shows the success of the auto-locate method, and how often the correct replacement for an errorful subpiece was in the n-best list (N=50).

Table 1: Base Recognition and Auto-Locate Statistics

Statistic	Experiment 1	Experiment 2
Word Accuracy	93.1%	88.6%
Sentence Accuracy	63.1%	46.7%
Auto-Locate Success	83.3%	90.6%
Correct hypo. in N-best	91.0%	83.7%

Table 2 shows the success rates for the various repair methods in both experiments. The column labeled "Highlight" reports the results when the errorful section was highlighted exactly by hand. The other column gives the results when the highlighting was with the auto-locate method described in section 2.1.

Table 2: Repair Method Success Rates

Repair Method	Highlight	Auto-Locate
Exp. 1: Respeak	70.1%	64.6%
Exp. 1: Spell	82.6%	70.8%
Exp. 1: Respeak+Spell	84.0%	73.6%
Exp. 2: Respeak	67.4%	62.7%

Table 3 shows the improvements in overall sentence accuracy when using the separate and combined repair mechanisms.

Table 3: Improvement of Sentence Accuracy

Sentence Accuracy vs. Repair Method	Hand Highlighting	Voice Highlighting
No Repair (baseline)	63.1%	63.1%
Respeak Repair	83.8%	83.1%
Spell Repair	88.5%	84.1%
Respeak+Spell Repair	89.0%	86.4%

5. CONCLUSION

These results are very encouraging, indicating that even these simple methods can be very effective. The assumption that the correct transcription for the subpiece must be in the n-best list somewhere definitely restricts the possible situations that can be repaired, although, for our recognizer, with 50 hypotheses in each n-best list, the correct subpiece was contained in the n-best list about 85% of the time overall. Increasing N would increase the likelihood of the correct alternative existing in the n-best list, and so should increase the repair success rates further.

Our results indicate that repeating or spelling a misrecognized subsection of an utterance can be an effective way of repairing more than two thirds of recognition errors. These techniques alone do not guarantee the ability to correct a misrecognition every time, but when used as important components of a total speech interface, these and similar improvements should lead to greater user acceptance of speech interfaces in practical applications.

6. FUTURE WORK

Many possible avenues of future research exist, all with the goal of making speech interfaces more effective and natural for users. Alternative information cues and sources should be sought out and used to assist in error location and repair. More methods should be found which give the user greater flexibility and choice in how to communicate repair information to a speech interface. User studies will indicate the most natural methods of speech repair, and will reiterate the need for robustness.

7. ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Advanced Research Projects Agency, and ATR Interpreting Telecommunications Research Labs of Japan.

8. REFERENCES

[1] C. Nakatani and J. Hirschberg. *A Speech-First Model for Repair Identification in Spoken Language Systems.* In Proceedings of the ARPA Workshop on Human Language Technology. March, 1993.

[2] O. Schmidbauer and J. Tebelskis. *An LVQ based Reference Model for Speaker-Adaptive Speech Recognition.* ICASSP 1992, volume 1, pages 441-444.

[3] H. Hild. *Speaker-Independent Connected Letter Recognition with a Multi-State Time Delay Neural Network.* 3rd European Conference on Speech, Communication and Technology (EUROSPEECH) 1993.