Modeling Spontaneous Speech Events During Recognition

Peter A. Heeman

Computer Science and Engineering
Oregon Graduate Institute of Science and Technology
20000 NW Walker Rd., Beaverton OR 97006
heeman@cse.ogi.edu

ABSTRACT

In spontaneous speech, speakers segment their speech into intonational phrases, and make repairs to what they are saying. However, techniques for understanding spontaneous speech tend to treat these events as noise, in the same manner as they handle out-of-grammar constructions and misrecognitions. In our approach, we advocate that these events should be explicitly modeled. We modify the speech recognition process so that it not only models determines the words that the user is saying, but also models intonational phrasing and speech repairs. This not only improves speech recognition performance but also results in a much richer output from the recognizer, with speech repairs resolved and intonational phrase boundaries identified.

1. INTRODUCTION

To enable spoken dialogue systems to advance towards more collaborative interaction between humans and computers, we need to deal with language as it is actually spoken. In natural speech, speakers group their words into intonational phrases and make repairs to what they are saying. Consider the following speaker's turn from the Trains corpus [Heeman and Allen, 1995].

Example **1 (d93-13.3 utt63)**
um it'll be there it'll get to Dansville at three a.m. and then you wanna do you take tho- want to take those back to Elmira so engine E two with three boxcars will be back in Elmira at six a.m. is that what you wanna do

From reading the word transcription, the reader should immediately notice the prevalence of *speech repairs*, where speakers go back and change or repeat something they just said. Fortunately for hearers, speech repairs tend to have a standard form. The *reparandum* is the stretch of speech that the speaker is replacing; it might end in the middle of a word, resulting in a word fragment. The end of the reparandum is called the *interruption point*. There can also be an editing term, consisting of fillers, such as `uh' and `um', or cue phrases, such as `let's see', `well', and `okay'. This is then followed by the *alteration*, which is the replacement for the reparandum.

Speech repairs are very prevalent in spontaneous speech. In the Trains corpus, 10% of all words are part of the editing term or reparandum of a speech repair, and 54% of all speaker turns with at least 10 words have at least one repair. To determine the speaker's intended message, speech repairs need to be *resolved*: they need to be *detected*, by finding their interruption point, and *corrected*, by determining the extent of their reparandum and editing term.

In addition to making repairs, speakers also break their turn of speaking into intonational phrases, which are signaled through variations in the pitch contour, phoneme lengthening and pauses. Previous research has shown that intonational information can reduce syntactic ambiguity for humans [Beach, 1991] and for computer parsers [Ostendorf et al., 1993]. Although intonational phrases might not be the ideal unit for modeling interaction in dialogue, it is undoubtedly a major component of any definition.

We refer to speech repairs and intonational phrasing as *spontaneous speech events*. We now reshow our earlier example annotated in terms of them. Reparanda are indicated in italic, with the alteration starting on a new line indented to start at the reparandum onset. Intonational phrase boundaries are marked with `**%'.**

Example **3 (d93-13.3 utt63)**
um *it'll be there*
    it'll get to Dansville at three a.m. **%**
and then *you wanna*
       do you *take tho-*
          want to take those back to Elmira **%**
so engine E two with three boxcars will be back in Elmira at six a.m. **%**
is that what you wanna do **%**

Although the spontaneous speech events of speech repairs and intonational phrasing are much more common in human-human speech than in human-computer speech [Oviatt, 1995]. In fact, one line of research work involves investigating means of structuring the interaction with the user so as to reduce the complexity of the user's speech, both in terms of the number of disfluencies and the linguistic and intonational structure of the speech [Oviatt, 1995]. However, as time advances, we will want to build human-computer interfaces that can collaborate with us on difficult problem-solving tasks [Allen et al., 1995]. It will become imperative to allow both the human and computer to be able to freely contribute to the dialogue, rather than having them focus on the form their interaction [Price, 1997]. Thus, we will need to allow for the full richness of spontaneous speech, with its apparent imperfections, such as speech repairs, and complex intonational phrasing.

To deal with these spontaneous speech events, it has become popular to use robust parsing techniques. For understanding spontaneous speech, speech repairs are not the only phenomena that create problems; one also needs to deal with word misrecognitions and out-of-grammar constructions. All three of these problems tend to be lumped together and given to a robust parser. Ward [1991] used a robust semantic parser to look for sequences of words that matched grammar fragments associated with slots of case frames. The parser tries to fill as many slots as possible. If a slot is only partially filled, it is abandoned. If a slot is filled more than once, the latter value is taken [Young and Matessa, 1991]. Others have adopted a similar approach: Rose and Lavie [2001] describe a robust parser, which incorporates a skipping mechanism, with a feature unification grammar; and van Noord describes using a skipping mechanism in parsing word graphs.

Rather than view spontaneous speech events as noise in the input to a robust parser, we take a different approach. We advocate that speech repairs and intonational phrasing should be explicitly modeled. There are local cues, such as editing terms, word correspondences, and pauses, which give evidence for these events. Hence, we should be able to automatically identify intonational phrases and resolve speech repairs. By modeling these events, we will have a richer understanding of the speech. This will simplify later syntactic and semantic processing, since such processing can start from enriched output rather than trying to cope with the apparent ill-formedness that spontaneous speech events cause [Core and Schubert, 1999]. This should also make it easier for these processes to deal with the other problems of understanding spontaneous speech: namely misrecognitions and out-of-grammar constructions.

Speech repairs and intonational phrasing are intertwined with the speech recognition problem of predicting the next word given the previous context [Heeman and Allen, 1999]. Hence, our approach is to redefine the speech recognition problem so that it includes the resolution of speech repairs and identification of intonational phrases. We also include the tasks of part-of-speech (POS) tagging and discourse marker identification, since these tasks are also intertwined with resolving speech repairs and identifying intonational phrasing. Since all tasks are being resolved in the same model, we can account for the interactions between the tasks in a framework that can compare alternative hypotheses for the speakers' turn. Not only does this allow us to model the spontaneous speech events, but it also results in an improved language model. Furthermore, speech repairs and phrase boundaries have acoustic correlates, such as pauses between words. By resolving speech repairs and identifying intonational phrases during speech recognition, these acoustic cues, which otherwise would be treated as noise, can give evidence as to the occurrence of these events, and further improve speech recognition results.

In the rest of the paper, we first give a brief overview of speech recognition language modeling. We then present a simplified version of our model of speech repairs and intonational phrases. We then give the results of running our model on the Trains corpus. Finally, we present the conclusions and future work.

2. SPEECH RECOGNITION

The goal of a speech recognizer is to find the sequence of words $\hat{W}$ that is maximal given the acoustic signal A.

$$\hat{W} = \arg\max_W \Pr(W \mid A) = \arg\max_W \frac{\Pr(AW)\Pr(W)}{\Pr(A)} = \arg\max_W \Pr(A \mid W)\Pr(W)$$

The above shows the speech recognition problem rewritten as the product of two probability distributions that need to be estimated: the first is the *acoustic model* $\Pr(A \mid W)$ and the second is the *language model* $\Pr(W)$. The language model probability can be expressed as the product of the conditional probability of each word given the words that precede it. This is shown below, where we rewrite the sequence W explicitly as the sequence of N words.

$$\Pr(W_{1,N}) = \prod_{i=1}^{N} \Pr(W_i \mid W_{1,i-1})$$

To estimate the probability distribution in the above line, a training corpus is used to determine the relative frequencies. Due to sparseness of data, one must define *equivalence classes* amongst the contexts $W_{1,i-1}$, which can be done by limiting the context, or by using decision trees.

3. AUGMENTING THE RECOGNIZER

The basic speech recognition model uses a simplistic language model, in which the probability of a word only takes into account the previous words. The occurrence of speech repairs and intonational boundaries, however, also affects which word will occur next. Consider the following example.

***Example* 4 (d93-3.2 utt 45)**

which engine   are we   are we taking

reparandum ip

After seeing the words "which engine are we," the probability of then seeing the word "are'" again would be very low. The word trigram `"are we are" rarely occurs. But, by hypothesizing a speech repair, its probability would be much higher. Furthermore, there are acoustic cues that can be used to give independent confirmation of the repair, thus further improving our ability to model these events.

To incorporate speech repair and intonational phrasing into our language model, we redefine the speech recognition problem so as to include extra variables that will be tagged for the occurrence of these events. Let I_i indicate whether word W_{i-1} ends an intonational phrase and let R_i indicate whether word W_{i-1} is the interruption point of a speech repair. The speech recognition problem can now be cast as finding the best interpretation for the repair and intonation variables along with the best word sequence.

$$\widehat{WRI} = \arg\max_W \Pr(WRI \mid A) = \arg\max_{WRI} \Pr(A \mid WRI)\Pr(WRI) \approx \arg\max_{WRI} \Pr(A \mid W)\Pr(WRI)$$

The second probability distribution is our new language model, which can be rewritten as follows.

$$\Pr(W_{1,N} R_{1,N} I_{1,N}) = \prod_{i=1}^{N} \Pr(I_i \mid W_{1,i-1} R_{1,i-1} I_{1,i-1})\Pr(R_i \mid W_{1,i-1} R_{1,i-1} I_{1,i})\Pr(W_i \mid W_{1,i-1} R_{1,i} I_{1,i})$$

Our full model, which is described elsewhere [Heeman and Allen, 1999], includes five additional variables. One variable is used to model the POS tag for each word, which allows us to model shallow syntactic knowledge and how it correlates with speech repairs and intonational phrasing. Another variable is used to model editing terms, which sometimes accompany speech repairs. The other three variables are used to correct the speech repair, which is determining the extent of the reparandum. Modeling the reparandum helps in detecting speech repairs and improves speech recognition performance. This improvement is because there are often strong word correspondences between the reparandum and alteration, which gives evidence for the repair, and helps predict the words in the alteration. Furthermore, the alteration tends to be a fluent continuation of the words before the reparandum, giving us further evidence of the repair and of the words of the alteration. We use a variable to indicate the reparandum onset, and for each word of the alteration, a variable indicates which word of the reparandum it corresponds to, and another variable indicates the type of correspondence, whether it is a word match, same POS tag, or other. This gives us a total of eight variables for our language model of spontaneous speech.

4. RESULTS

To test our model, we ran experiments on the Trains corpus using a six-fold cross validation procedure. We tested three versions three different language models [Heeman, 1999]. The first is a traditional word-based language model, the second incorporated modeling POS tags, and the third is our full model of speech repairs and intonational phrasing (and POS tags). The results are given in Table 1. We report the results in terms of perplexity and word error rate. *Perplexity* is a measure of how well the probability distributions of the language model predict the data in a test set $w_{1,N}$, and is calculated as 2^H, where $H = -\dfrac{1}{N}\sum_{i=1}^{N}\log_2 \widehat{\Pr}(w_i \mid w_{1,i-1})$. Lower perplexity values

indicate improvements in predicting the test data. *Word error rate* measures the percentage of mistakes that a speech recognizer makes using the language model. We used a large vocabulary continuous speech recognizer that has completed in the Broadcast News task developed at OGI.

	Perplexity	WER
Word – Based Model	24.8	26.0
POS – Based Model	22.6	24.9
Full Model	21.3	24.6

Table 1: Impact on Speech Recognition

We see that by incorporating the additional modeling, we are able to improve the speech recognizer's performance by 5.4%, as measured with word error rate. This model only makes limited use of acoustical cues, namely pauses between words. By using richer acoustic cues of speech repairs and intonational phrases, we should be able to further improve the results (cf. [Stolcke et al., 1999]).

We have also run tests to determine how well we can identify intonational phrase boundaries and speech repairs [Heeman and Allen, 1999]. We again used a six-fold cross validation procedure. We report results in terms of *recall* and *precision*. The recall rate is the number of times that the algorithm correctly identifies an event over the total number of times that it actually occurred. The precision rate is the number of times the algorithm correctly identifies it over the total number of times it identifies it.

For intonational phrase boundaries, we distinguish between those that occur within a speaker's turn from those that occur at the end. This is because our model uses end-of-turn information as part of its input, and since almost all turns end with an intonational phrase boundary, it easily learns this regularity. As for the within turn boundaries, the model achieves a recall rate of 71.8% with a precision of 70.8%.

For detecting speech repairs, we achieved a recall rate of 76.8% with a precision of 86.7%. Although our model classifies repairs into three different types (fresh start, modification, and abridged), we count a repair as correct as long as its interruption point was identified, without regard to whether its type was correctly determined. Furthermore, when multiple repairs have contiguous reparanda, we count all repairs involved (of the hand-annotations) as correct as long as the combined reparandum is correctly identified. For correcting speech repairs, we achieved a recall rate of 65.9% and a precision of 74.3%. A repair is counted as correctly corrected if it was identified and the extent of the reparandum was correctly determined.

5. CONCLUSION AND FUTURE WORK

Previous work in spoken language interfaces has regarded speech repairs and intonational phrasing as noise in the dialogue, which needs to be skipped over and not understood (e.g. [Ward, 1991]). Others have investigated means of structuring the user's interactions so as to reduce the complexity of the user's speech, thus making it easier to understand [Oviatt, 1995]. However, as time advances, we will want to build human-computer interfaces that can collaborate with users on difficult problem-solving tasks [Allen et al., 1995]. It will become imperative to allow both the human and computer to be able to freely contribute to the dialogue, using the full richness of language, with its apparent imperfections, such as speech repairs and complex intonational phrasing.

In this paper, we reported on our work in which we model these spontaneous speech events. We redefined the speech recognition task so as to also include identifying intonational phrases and resolving speech repairs. This allows us to better account for the words involved in a speaker's turn, as evidenced by the improved word recognition rates. It also allows us to return a more meaningful analysis of the speaker's turn, with speech repairs and intonational phrases identified. This richer output should make it easier for later processing to deal with spontaneous speech. Hence, interfaces of the future will be able to better deal with the occurrence of spontaneous speech events. In fact, human-computer interfaces might when a user makes a speech repair, this gives evidence that the speaker might be uncertain about what he is saying. The human-computer interface should then modify its response appropriately, for instance by double checking the information.

Much work remains to be done. One area of research that we are pursuing is incorporating higher level syntactic and semantic processing. This would not only allow us to give a much richer output from the model, but it would also allow us to account for interactions between this higher level knowledge and modeling speakers' utterances, especially in detecting the ill-formedness that often occur with speech repairs. It would also help in finding richer correspondences between the reparandum and alteration, such as between the noun phrase and pronoun in the following example.

***Example* 5 (d93-14.3 utt27)**

the engine can take as many | um | it can take up to three
reparandum ↑ et alteration
ip

A second area of research is to incorporate better acoustical cues. We currently only exploit inter word pauses. This is a rich source of information for detecting (and distinguishing between) intonational phrases and interruption points of speech repairs. It would also help in determining the reparandum onset [Nakatani and Hirschberg, 1994], which tend to occur at intonational boundaries. Acoustic modeling is also need to identify word fragments, which accompany many speech repairs. As we further exploit acoustic cues, this will improve our ability to detect speech repairs and intonational phrases, which will have the added benefit of improving our speech recognition results.

7. ACKNOWLEDGMENTS

8. REFERENCES

Allen, J., Schubert, L., Ferguson, G., Heeman, P., Hwang, C. H., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M., and Traum, D. R. (1995). The Trains project: A case study in building a conversational planning agent. *Journal of Experimental and Theoretical AI*, 7:7-48.

Beach, C. M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language*, 30(6):644-663.

Core, M. and Schubert, L. (1999). A syntactic framework for speech repairs and other disruptions. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, College Park, Maryland.

Heeman, P. A. (1999). Modeling speech repairs and intonational phrasing to improve speech recognition. In *Automatic Speech Recognition and Understanding Workshop*, Keystone Colorado.

Heeman, P. A. and Allen, J. F. (1995). The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium.

Heeman, P. A. and Allen, J. F. (1999). Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialog. *Computational Linguistics*, 25(4):527--572.

Nakatani, C. H. and Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, 95(3):1603-1616.

Ostendorf, M., Wightman, C., and Veilleux, N. (1993).Parse scoring with prosodic information: an analysis / synthesis approach. *Computer Speech and Language*, 7(2):193--210.

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9:19-35.

Price, P. (1997). Spoken language understanding. In Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A., and Zue, V., editors, *Survey of the State of the Art in Human Language Technology*. Cambridge University Press.

Rosé, C. P. and Lavie, A. (2001). Balancing robustness and efficiency in unification-augmented context-free parsers for large practical applications. In *Robustness in Language and Speech Technology*. Kluwer Academic Publishers.

Stolcke, A., Shriberg, E., Hakkani-Tür, D., and Tür, G. (1999). Modeling the prosody of hidden events for improved word recognition. In *Proceedings of the 6th European Conference on Speech Communication and Technology*.

van Noord, G. (2001). Robust parsing of word graphs. In *Robustness in Language and Speech Technology*. Kluwer Academic Publishers.

Ward, W. (1991). Understanding spontaneous speech: The Phoenix system. In *Proceedings of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, pages 365-367.

Young, S. R. and Matessa, M. (1991). Using pragmatic and semantic knowledge to correct parsing of spoken language utterances. In *Proceedings of the 2nd European Conference on Speech Communication and Technology*, pages 223-227, Genova, Italy.