

# MODELING SPEECH REPAIRS AND INTONATIONAL PHRASING TO IMPROVE SPEECH RECOGNITION

Peter A. Heeman

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

## ABSTRACT

The spontaneous speech events of speech repairs and intonational phrasing cause disruptions in the local context, and this disruption prevents traditional language models from being able to properly predict the words in the vicinity of these events. The solution is to use a language model that can account for these spontaneous speech events. In this paper, we use such a model to rescore word graphs. This gives a small but significant decrease in the word error rate of 1.2%, in addition to an improvement of 4.4% from modeling the syntactic role of the words. Furthermore, as modeling of spontaneous speech events improves, word recognition results should also improve.

## 1. INTRODUCTION

To enable spoken dialogue systems to advance towards more collaborative interaction, systems need to handle language as it is actually spoken. People not only utter a string of words, but they group them 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 (Levelt, 1983). 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', and cue phrases, such as 'let's see', 'well', and 'okay'. This is then followed by the *alteration*, which is the replacement for the reparandum. Below, we illustrate this analysis on the first repair from the above speaker turn.

*Example 2 (Repair)*

um it'll be there ↑ it'll get to Dansville at three a.m.  
reparandum *ip* alteration

In the Trains corpus, 54% of turns with at least 10 words have a repair, 10% of all words are part of the editing term or reparandum of

a speech repair. In order to understand spontaneous speech, the extent of the reparanda and editing terms must be determined, which we refer to as *correcting* the repair. Hearers seem to understand such speech effortlessly. In fact, they have difficulty recalling the location of the repair (Martin and Strange, 1968) and the words in the reparanda (Bard and Lickley, 1997). So, hearers must be detecting and correcting (determining the extent of the reparandum and editing term) repairs very early in processing the speech.

In addition to making repairs, speakers also break their turns into intonational phrases, which are signaled through variations in the pitch contour, segmental lengthening and pauses. Previous research has shown that intonational information can reduce syntactic ambiguity for humans (Beach, 1991) and for computer parsers (Ostendorf, Wightman, and Veilleux, 1993). Other researchers have proposed segmenting speech into speech acts (Mast et al., 1996) or into linguistic clauses (i.e., Meteer and Iyer, 1996; Stolcke et al., 1999). There is no clear consensus as to the right approach, however. Although intonational phrases might not be the ideal unit for modeling interaction in dialogue, it does captures speaker intention and is a major component of any definition (Traum and Heeman, 1997).

Now that we have introduced the spontaneous speech events, we show our example annotated in terms of them. Repair 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 %

In order to understanding spontaneous speech, we must identify the intonational phrases and detect and correct the speech repairs. Hearers seem to be able to do these tasks very early in processing the speech, and hence there must be enough cues in the local context to make this feasible. In previous work (Heeman and Allen, 1999), we expanded the speech recognition task to account for spontaneous speech events: in addition to finding the best word sequence given the acoustic signal, we also jointly find the best POS tags (syntactic category of each word), speech repair, intonational phrase and discourse marker interpretation. This is done by adding extra variables into the speech recognition process to

account for the spontaneous speech events (see Section 4). Since all tasks are being resolved together in the same model, we can account for the interactions between the tasks because the framework allows alternative hypothesis about the speaker’s turn to be compared. Our approach of modeling spontaneous speech events as part of the speech recognition process allows us to use acoustic cues, such as pauses and syllable lengthening, which otherwise would be treated as noise, to give evidence as to the occurrence of these events. This further improves modeling of the spontaneous speech events, which should improve the speech recognition word-error rate. We have shown that our model reduces the perplexity on a test corpus (Heeman and Allen, 1999), and we have shown that a simpler version of our model, which only adds POS tags, results in a 4.4% decrease in word error rate (Heeman, 1999).

In this paper, we will show that our spontaneous speech language model further improves the word error rate. Furthermore, by modeling spontaneous speech events, we will have a richer understanding of the speech, with intonational phrases and speech repairs annotated. This will simplify later syntactic and semantic processing, since such processing can start from our enriched output rather than trying to cope with the apparent ill-formedness of spontaneous speech. This will also make it easier for these processes to cope with the added syntactic and semantic variance that spontaneous speech seems to license.

In the rest of the paper, we first argue that there are interactions between modeling spontaneous speech events and speech recognition. Traditional language models do not model these interactions and hence have difficulties recognizing the words in the vicinity of these events. We then show that the word graph does not overly penalize such words (with the exception of word fragments), and hence is suitable for rescoring. Next, we give a brief overview of our statistical language model. We give the results of applying this model in rescoring word graphs. We then compare our approach with others that have been proposed, and then conclude.

## 2. INTERACTIONS WITH SPEECH RECOGNITION

The tasks of identifying intonational phrases and detecting and correcting speech repairs have strong interactions with the speech recognition task of predicting the next word. The interruption point of speech repairs and the boundaries of intonational phrases create disruptions in the local context, making it more difficult to predict the next word. Furthermore, many speech repairs have word correspondences across the interruption point, and these correspondences should simplify word prediction. Conversely, the presence of a disruption and word correspondences can provide evidence that a repair occurred. Hence, speech recognition must be done simultaneously with modeling spontaneous speech events. If the interactions are not modeled, the speech recognizer will have problems predicting the words in the vicinity of these events.

To illustrate this problem, we ran our speech recognizer (Wu et al., 1999) with a language model built with the CMU toolkit (Rosenfeld, 1995). The language model was trained on our spontaneous speech corpus, but does not model spontaneous speech events. The first bar of each pair in Figure 1 gives the percentage of words that were recognized (misses and substitutions) in different vicinities of the turn (we exclude turns with word fragments). From left to right, the vicinities are words immediately before the interruption point of speech repairs, editing term words, words that immediately follow the interruption point or editing term, words that immediately precede an intonational phrase boundary (not in-

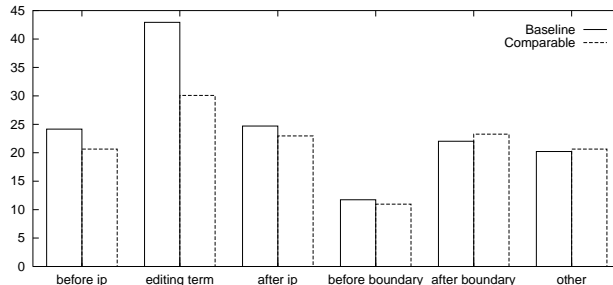


Figure 1: Word-error rates from traditional language model

cluding end-of-turn words), words that immediately follow an intonational phrase break and all other words. We see that, with the exception of the words before intonational boundaries, the word error rate is worse for words in the vicinity of a spontaneous speech event than for other words.

The above analysis, however, is very crude because it does not take into account how spontaneous speech events interact with word distributions. For instance, the word immediately before an intonational boundary is likely to be a content word, while the word after it is likely to be a function word. Function words tend to be more poorly recognized than content words, which is the main reason why words after intonational phrase boundary have a lower recognition rate than words before. To give a fairer comparison, the second bar of each pair shows the overall word error rate but weighed to have the same POS distribution as the category. By comparing the first and second bars of each pair, we see that, with the exception of words immediately after the intonational phrase boundary, words in the vicinity of the spontaneous speech events are not as well recognized. For instance, the word before the interruption of a speech repair is not recognized 24.2%, whereas the average misrecognition for similar words is 20.7%; editing terms are misrecognized at 42.9%, rather than the expected rate of 30.1%; the first word after the repair is misrecognized at 24.7%, rather than 23.0%; and the word before the intonational phrase boundary is misrecognized at 11.7%, rather than 11.0%.

## 3. ADEQUACY OF THE WORD-GRAPH

Word graphs are an effective way of interfacing a speech recognizer with more complex language modeling (Johnson, Harper, and Jamieson, 1998). In this section, we show that they can be used for rescoring with a spontaneous speech language model. For turns with no word fragments, we found the path through the word

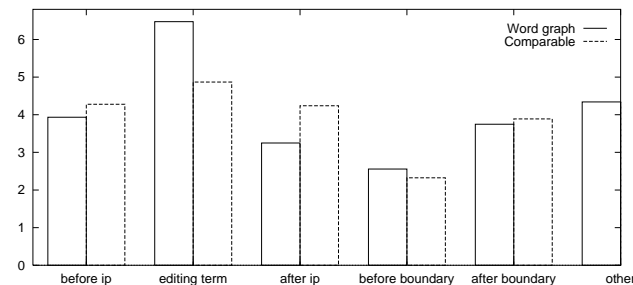


Figure 2: Word error rate for optimal path through graph

graph with the lowest word error rate. We aligned this path with the correct transcription and then did the same comparison as we did for Figure 1. By comparing the first and second bars of each pair, we see that words in the editing term and before intonational phrase endings are not as well recognized as the average. Editing terms have a word-graph misrecognition rate of 6.5%, whereas the average misrecognition rate for similar words is 4.9%; and the word before the intonational boundary has a misrecognition rate of 2.6%, rather than 2.3%. However, the word before the repair has a misrecognition rate of 3.9%, rather than 4.3%; the word after the repair has a rate of 3.3%, rather than 4.2%; and the word after the intonational phrase has a rate of 3.8%, rather than 3.9%. So, the word graph does not overly penalize words in the vicinity of spontaneous speech events, and hence should prove suitable as an interface for modeling spontaneous speech events. Word fragments remain problematic, which occur on 26% of all repairs.

#### 4. MODELING SPONTANEOUS SPEECH EVENTS

To model spontaneous speech events, we enlarge the speech recognition problem (Heeman and Allen, 1999). Rather than trying to find the best word sequence for the acoustic signal, we also try to find the best POS tags for the words, and the best phrase boundaries and speech repair interpretation. This is done by adding extra variables into the speech recognition equations. We add the variable  $P_i$  to indicate the POS tag for word  $i$ ,  $I_i$  to indicate if word  $i - 1$  ends an intonational phrase,  $R_i$  to indicate if word  $i - 1$  is the interruption point of a repair,  $E_i$  to indicate the extent of the editing term, and three other variables for correcting speech repairs, which for expository purposes we group together as  $C_i$ . The speech repair variables capture many different sources of information that can be used for modeling spontaneous speech events. They let us model the effect of the disruption in the context, fluent continuations between the first word of the alteration and the speech that precedes the reparandum, and word correspondences between the reparandum and alteration. The new speech recognition equation is as follows.

$$\begin{aligned} \hat{W}\hat{P}\hat{C}\hat{R}\hat{E}\hat{I} &= \arg \max_{WPCREI} \Pr(WPCREI|A) \\ &= \arg \max_{WPCREI} \Pr(A|WPCREI) \Pr(WPCREI) \end{aligned}$$

The  $\Pr(A|WPCREI)$  is the acoustic model, and  $\Pr(WPCREI)$  is the language model. We can rewrite the language model term as

$$\Pr(W_{1,N} P_{1,N} C_{1,N} R_{1,N} E_{1,N} I_{1,N})$$

where  $N$  is the number of words in the sequence. We now rewrite this term as follows.

$$\begin{aligned} &\Pr(W_{1,N} P_{1,N} C_{1,N} R_{1,N} E_{1,N} I_{1,N}) \\ &= \prod_{i=1}^N \Pr(W_i P_i C_i R_i E_i I_i | W_{1,i-1} P_{1,i-1} C_{1,i-1} R_{1,i-1} E_{1,i-1} I_{1,i-1}) \\ &= \prod_{i=1}^N \Pr(I_i | W_{1,i-1} P_{1,i-1} C_{1,i-1} R_{1,i-1} E_{1,i-1} I_{1,i-1}) \\ &\quad \Pr(E_i | W_{1,i-1} P_{1,i-1} C_{1,i-1} R_{1,i-1} E_{1,i-1} I_{1,i-1}) \\ &\quad \Pr(R_i | W_{1,i-1} P_{1,i-1} C_{1,i-1} R_{1,i-1} E_{1,i-1} I_{1,i-1}) \\ &\quad \Pr(C_i | W_{1,i-1} P_{1,i-1} C_{1,i-1} R_{1,i-1} E_{1,i-1} I_{1,i-1}) \\ &\quad \Pr(P_i | W_{1,i-1} P_{1,i-1} C_{1,i-1} R_{1,i-1} E_{1,i-1} I_{1,i-1}) \\ &\quad \Pr(W_i | W_{1,i-1} P_{1,i-1} C_{1,i-1} R_{1,i-1} E_{1,i-1} I_{1,i-1}) \end{aligned}$$

The language model has eight probability distributions, each with a very rich context. We use a decision tree learning algorithm to divide the context into a hierarchy of equivalence classes and use interpolated estimation on the hierarchy to estimate the probability distributions (Bahl et al., 1989). In order to help the decision tree algorithm, we transform the context into a simpler representation that allows it to ask more meaningful questions. We also adjust the probability distributions to take into account the presence of pauses (Heeman and Allen, 1999).

#### 5. RESULTS

Table 1 gives the results of rescored word graphs produced by our recognizer using the baseline language model trained on our spontaneous speech corpus. A six-fold cross-validation procedure. All turns were included, including those with word fragments. We computed both the perplexity (of the known words) that each model assigned to the test corpus, and the word error rate. The first two rows compare the baseline model against a version of our model that accounts for POS tags, but not the spontaneous speech events. Incorporating POS tags results in a 4.2% improvement in word error rate and an 8.9% improvement in perplexity (Heeman, 1999). The third row gives the results of using our spontaneous speech language model. This model further improves the word error rate by 1.2% and improves the perplexity by 5.8%. The improvement in word error rate was judged to be significant by the Wilcoxon test on the 34 speakers in the corpus.

	Perplexity	WER
Word-Based Model	24.8	26.0
POS-Based Model	22.6	24.9
Full Model	21.3	24.6
Intonational Oracle	20.3	—
Repair Oracle	19.3	—
Intonation & Repair Oracle	18.3	—

Table 1: Comparison of Different Models

Figure 3 compares the performance of the three models in the vicinity of the spontaneous speech events (for turns without word fragments). With the exception of words before the intonational phrase boundary, the full model outperforms the POS-based model. The model even gives a small improvement to words that

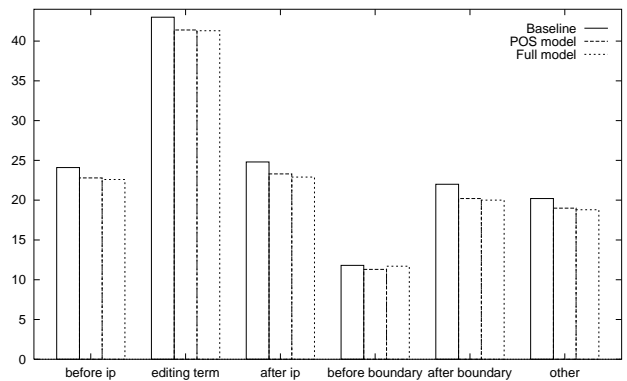


Figure 3: Comparison of Different Models

are not in the direct vicinity of the spontaneous speech events. This is because the training data for the full model distinguishes the spontaneous speech events from the rest, and hence the model of the rest of the speech is more precise.

## 6. POTENTIAL IMPACT

To further show the value of modeling speech repairs and intonational phrase boundaries, we ran a set of cheating experiments in which we assume perfect intonational phrase boundary detection and perfect detection of the interruption point of speech repairs and the extent of their editing terms. Speech repair correction is still handled by our statistical model. The results are given in the last three rows of Table 1. The third last row gives the results using the intonational boundary oracle, the second last row gives the results using the oracle for detecting speech repairs and their editing terms, and the last column combines both oracles. As can be seen, better detection of both speech repairs and intonational phrase boundaries leads to decreases in perplexity. Perfect detection of both would lead to a further 14.0% reduction in perplexity, giving an overall improvement of 19.0% over the POS-based model and 26.2% over the word-based model. These results illustrate the potential improvement that can be gained in language modeling from improving the modeling of speech repairs and intonational phrase boundaries. Part of this improvement can be obtained by improving the acoustic cues used to detect the spontaneous speech events (cf. Stolcke *et al.*, 1999), and part due to improving the probability estimates for these events.

## 7. RELATIONSHIP TO OTHER WORK

Although substantial research has been done in the area of intonational phrasing and speech repairs, very little work has been done on incorporating this research work with speech recognition. One of the only exceptions has been the work of Shriberg and Stolcke (Stolcke *et al.*, 1999). In the latter work, they rescored an n-best list using a ‘hidden event’ language model, in which they sum over different types of spontaneous speech events in computing the probability of the next word. The events are single word repetition, single word deletion, filler, other repairs, and sentence boundaries. They also incorporate prosodic cues. With this approach, they were able to show that modeling hidden events resulted in an improvement in word-error rate from 47.9% for a traditional language model on the Switchboard corpus to 47.0%. Without the acoustic component, their word-error rate is 47.6. Our work differs from theirs in that we use a more complex model that can account for the interaction between speech repair detection, speech repair correction, intonational phrase boundaries and syntactic analysis (Heeman and Allen, 1999).

## 8. CONCLUSION

In this paper, we argued that the speech recognition task needs to incorporate modeling of spontaneous speech events. With the exception of word fragments, this incorporation can be done by rescored word graphs using a language model that accounts for spontaneous speech events. Although our spontaneous speech language model only gives a 1.2% improvement in word error rate, this is on top of a 4.4% improvement from incorporating POS tags. Our experiments in which we assume perfect modeling of spontaneous speech events show that as we further improve the modeling

of the spontaneous speech events, we should see further reductions in word error rate.

## 9. ACKNOWLEDGMENTS

We wish to thank James Allen, Chaojun Liu, Xintian Wu, and Yonghong Yan. This work was partially funded by the Intel Research Council.

## 10. REFERENCES

- Bahl, L. R., P. F. Brown, P. V. de Souza, and R. L. Mercer. 1989. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1001–1008.
- Bard, E. G. and R. J. Lickley. 1997. On not remembering disfluencies. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, pages 2855–2858.
- Beach, C. 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.
1999. POS tags and decision trees for language modeling. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 129–137.
- Heeman, P. A. and J. F. Allen. 1995. The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium.
- Heeman, P. A. and J. F. Allen. 1999. Speech repairs, intonational phrases and discourse markers: Modeling speakers’ utterances in spoken dialog. *Computational Linguistics*, 25(4).
- Johnson, M. T., M. P. Harper, and L. H. Jamieson. 1998. Effectiveness of word graphs for interfacing speech recognition and language models. In *Proceedings of the 5th International Conference on Spoken Language Processing*.
- Levelt, W. J. M. 1983. Monitoring and self-repair in speech. *Cognition*, 14:41–104.
- Martin, J. and W. Strange. 1968. The perception of hesitation in spontaneous speech. *Perception and Psychophysics*, 53:1–15.
- Mast, M., R. Kompe, S. Harbeck, A. Kiebling, H. Niemann, E. Nöth, E. G. Schukat-Talamazzini, and V. Warnke. 1996. Dialog act classification with the help of prosody. In *Proceedings of the 4th International Conference on Spoken Language Processing*, pages 1728–1731.
- Meteor, M. and R. Iyer. 1996. Modeling conversational speech for speech recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Ostendorf, M., C. Wightman, and N. Veilleux. 1993. Parse scoring with prosodic information: an analysis/synthesis approach. *Computer Speech and Language*, 7(2).
- Rosenfeld, R. 1995. The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation. In *Proceedings of the ARPA Spoken Language Systems Technology Workshop*.
- Stolcke, A., E. Shriberg, D. Hakkani-Tür, and G. Tür. 1999. Modeling the prosody of hidden events for improved word recognition. In *Proceedings of the 6th European Conference on Speech Communication and Technology*.
- Traum, D. R. and P. A. Heeman. 1997. Utterance units in spoken dialogue. In E. Maier, M. Mast, and S. LuperFoy, editors, *Dialogue Processing in Spoken Language Systems*, Lecture Notes in Artificial Intelligence. Springer-Verlag, pages 125–140.
- Wu, X., C. Liu, Y. Yan, D. Kim, S. Cameron, and R. Parr. 1999. The 1998 ogi-fonix broadcast news transcription system. In *DARPA Broadcast News Workshop*.