TR-IT-0148

Using Acoustic Clues to Distinguish Speech Repairs and Intonational Phrase Boundaries

Peter A. Heeman
ピーター ヒーマン

1996.1

In spoken dialogue, hearers can make use of acoustic clues to understand what the speaker is saying. Previous work has shown their utility in signaling intonational phrase boundaries and prominence in professionally read speech (Wightman and Campbell, 1994). However, in spoken dialogue, intonational phrase boundaries and speech repairs seem to share acoustic as well as lexical clues. So it is necessary to resolve both issues at the same time. In this paper, we extend the existing approach to account for both intonational phrase boundaries and speech repairs in spontaneous speech. We also incorporate these clues into a statistical model, based on a part-of-speech tagger, that takes into account lexical clues.

1 Introduction

In speech, hearers can make use of acoustic clues in understanding what the speaker is saying. Previous work has shown their utility in signaling intonational phrase boundaries and prominence in professionally read speech (Wightman and Campbell, 1994). However, spontaneous speech differs dramatically from read speech. First, conversants in spoken dialogue are not able to extend their full attention to the task of generating their utterances. The result is that their intonational clues might not be as robust as in professionally read speech.

Second, conversants must co-ordinate their efforts in managing the turn-taking activities (Sacks, Schegloff, and Jefferson, 1974). Conversants use clues to signal that they are done with the turn, one of these clues being intonational completeness (Ford and Thompson, 1991). Hence speakers might alter the lengths of pauses, locations of breaths, and use of filled pauses in order to maintain the turn.

Third, conversants are engaged in a collaborative activity (Clark and Wilkes-Gibbs, 1986), in which they are trying to minimize collaborative activity. This might license a hearer to interrupt the speaker, or for the speaker to not finish utterances if she thinks her point is understood.

Fourth, and most important to this work. is that spoken dialogue has an abundance of speech repairs. Speech repairs are dysfluencies where some of the words that the speaker utters need to be removed in order to correctly understand the speaker's meaning. As illustrated in the example below from the TRAINS corpus (d92a-5.2 utt34), they can be divided into three intervals, or stretches of speech: the *reparandum*, the *editing terms*, and the *alteration*.

we'll pick up a tank of uh the tanker of oranges

reparandum | *editing terms* *alteration*

interruption point

The reparandum is the stretch of speech that the speaker intends to replace, and this could end with a *word fragment*, where the speaker interrupts herself during the middle of the current word. The end of the reparandum is called the *interruption point* and is often accompanied by a disruption in the intonational contour. This is then followed by the editing terms, which can either be a filled pause, such as "um" or "uh" or a cue phrase, such as "I mean", "well", or "let's see". The last part is the alteration, which is the speech that the speaker intends as the replacement for the reparandum. In order to *correct* a speech repair, the reparandum and the editing terms need to be deleted in order to determine what the speaker intends to say.

Detection of speech repairs relies on the combination of a number clues, both acoustic and lexical. Experiments by Lickley and Bard (1992) have found that hearers were able to recognize a disfluency before they recognized the first word in the alteration in 66.5% of the cases. This suggests that acoustic clues can be used to good effect to recognize speech repairs. The problem is that some of the acoustic clues that signal speech repairs

are the same clues that signal intonational phrase boundaries. Pause durations have been used for intonational phrase detection (Wightman and Ostendorf, 1994) and have been reported as a significant clue for detecting speech repairs (i.e. (Nakatani and Hirschberg, 1993)). Preboundary lengthening was also used by Wightman and Ostendorf, expressed as the duration of the rhyme, the stretch of speech beginning of the last vowel to the end of the word. However, Clark (Clark, 1996) claims that syllable elongation and non-reduction of vowels often mark speech repairs.

In this paper, we use acoustic clues to account for both intonational phrase boundaries and speech repairs in spontaneous speech. Since the acoustic clues are too weak on their own to fully detect these events, we incorporate them into a statistical model, based on a part-of-speech tagger (Heeman and Allen, 1996), that takes into account lexical clues.

2 Classification of Transition Types

We classify transitions between two adjacent words, w_i and w_{i+1} with a variable W_{i+1}. Before we give the possible values, we introduce our speech repair classification.

As in (Heeman and Allen, 1996), we divide speech repairs into three groups, which are based on what the hearer needs to do to correct a speech repair. We divide speech repairs into three types: *fresh starts*, *modification repairs*, and *abridged repairs.* A fresh start is where the speaker abandons the current utterance and starts again, where the abandonment seems acoustically signaled (d93-12.1 utt30).

so it'll take um so you want to do what
reparandum ↑ *editing term* *alteration*
interruption point

The second type of repairs are the modification repairs. These include all other repairs in which the reparandum is not empty (d92a-1.3 utt65).

so that will total will take seven hours to do that
reparandum ↑ *alteration*
interruption point

The third type of repairs are the abridged repairs, which consist solely of an editing term (d93-14.3 utt42).

we need to um manage to get the bananas
↑ *editing term*
interruption point

Problematic to the above classification scheme are the repairs whose reparandum consists solely of a word fragment. Under this scheme, these will either be fresh starts or modification repairs. However, our input is the word transcription (as would be provided by an ideal speech recognizer), with the word fragments marked. Given this, a word fragment is much

like a filled pause in that we know we must always remove them. So, we will treat such repairs in the same class as the abridged repairs.

Given the above classification of speech repairs, we are now ready to define the transition type W_{i+1} between word w_i and w_{i+1}.

$W_{i+1} = \mathcal{C}$: w_i is followed by the interruption point of a fresh start and w_i is the onset of the alteration.

$W_{i+1} = \mathcal{M}$: w_i is followed by the interruption point of a modification repair and w_{i+1} is the onset of the alteration.

$W_{i+1} = \mathcal{P}ush$: w_{i+1} is the first word of an editing term.

$W_{i+1} = \mathcal{P}op$: w_{i+1} is the last word of an editing term.

$W_{i+1} = \mathcal{B}$: w_i is the end of an intonational phrase and is not followed by an interruption point, and neither w_i nor w_{i+1} is an editing term.

$W_{i+1} = \mathcal{P}$: Others.

To simplify our analysis, we ignore transitions to and from word fragments and filled pauses and we will not be dealing with the Push and Pop transitions. Hence, we will only be considering speech repairs that do have editing terms or word fragments.

3 The Trains Corpus

As part of the TRAINS project (Allen et al., 1995), which is a long term research project to build a conversationally proficient planning assistant, we have collected a corpus of problem solving dialogs (Heeman and Allen, 1995). The dialogs involve two human participants, one who is playing the role of a user and has a certain task to accomplish, and another who is playing the role of the system by acting as a planning assistant. The collection methodology was designed to make the setting as close to human-computer interaction as possible, but was not a *wizard* scenario, where one person pretends to be a computer. Rather, the user knows that he is talking to another person.

The corpus consists of 98 dialogs totaling six and a half hours in length and containing about 55,000 words, 5900 speaker turns, and 34 different speakers. These dialogs have been segmented into single speaker utterance files and word annotated using the Waves software (Ent, 1993). The corpus is available from the Linguistics Data Consortium on CD-ROM (Heeman and Allen, 1995).

The speech repairs in the dialog corpus have been hand-annotated. There is typically a correspondence between the reparandum and the alteration, and following (Bear et al., 1993), we annotate this using the labels m for word matching and r for word replacements (words of the same syntactic category). Each pair is given a unique index. Other words in

the reparandum and alteration are annotated with an x. Also, editing terms (filled pauses and clue words) are labeled with **et**, and the interruption point with **ip**, which will occur before any editing terms associated with the repair, and after the fragment, if present. The interruption point can also be marked as to whether the repair is a fresh start or a modification repair, in which cases, we use **ip:can** and **ip:mod**, respectively. The example below illustrates how a repair is annotated in this scheme.

engine two from Elmi- or engine three from Elmira
m1 r2 m3 m4 et m1 r2 m3 m4
ip:mod

Further details of this scheme can be found in (Heeman and Allen, 1996).

4 Acoustic Clues

In this section, we examine the predictive power of some acoustic clues. The use of most of these, has been motivated by the work of Wightman and Campbell (1994) who looked at detecting intonational phrase boundaries and prominence in read speech. Some of the clues are also motivated by descriptive work in understanding speech repairs.

In looking for evidence of acoustic correlates of word transition types, we will use the cumulative distributions for each clue. Note that individual distributions can be miss-leading because they only show the predictive power of the clue when used in isolation. Some clues might have interactions with another clue such that the two clues together are more predictive then each is separately. However, the distributions for each clue will give a rough indication of how useful the clue is.

Cumulative distributions show the range over which the feature takes its values. If for a given feature, the distributions for each transition type take on their values over different ranges, then the feature is able to distinguish the transitions. However, as we will see, there is no single feature that is able to separate the transitions. Rather each clue gives some evidence as to the transition type.

For the distributions, it will probably be the case that the curves for modification repairs and fresh starts are not as smooth (and as significant) as the ones for intonational phrase boundaries and plain transitions. This is due to the number of samples. We have 40000 plain transitions, 4800 boundary tones, 600 modification repairs, and 300 fresh starts.

Also note that the distributions were collected across many different speakers, both male and female. We do not have enough data on individual speakers to do speaker dependent analysis, and so when possible, we will try to account for speaker variation by normalizing the data.

4.1 Pausal Duration

The length of silences between words has been used by a number of researchers in studying intonational phrase boundaries (Wightman and Campbell, 1994) and in studying speech repairs (Nakatani and Hirschberg, 1993; O'Shaughnessy, 1992; Heeman and Allen, 1996). Its popularity is undoubtably due to how simple it is to calculate, and that it can be automatically derived from a word aligner. As with Heeman and Allen (1996), we use pausal durations as evidence of the transition type. Figure 1 gives the cumulative distribution for each transition type.

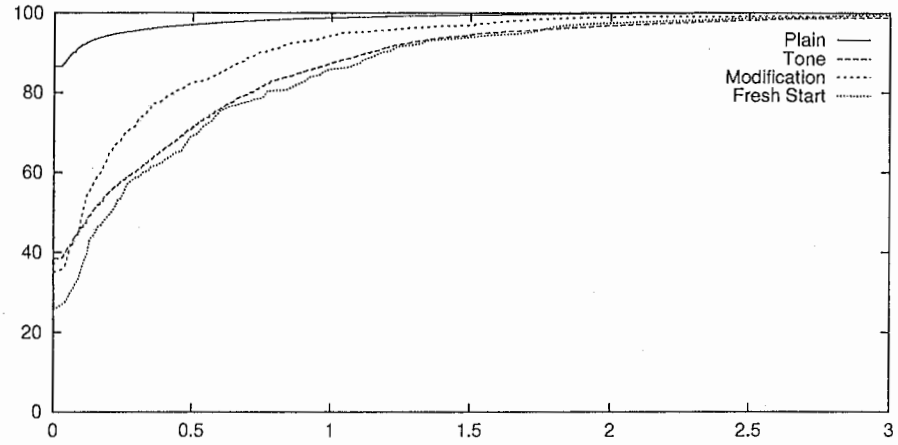

Figure 1: Pause lengths

As can be seen by the cumulative distribution curves, plain transitions tend to have smaller pauses; in fact about 85% of the plain transitions were found by the automatic aligner not to have any pause at all. The figure also shows how that fresh starts and modification repairs tend to have the longest pauses.

4.2 Durational Effects

Another important correlate of tone transitions is the lengthening that is often present in the word that precedes it (the word marking with the boundary tone). In fact, Wightman *et al* (1992) found that the lengthening is restricted to the rhyme of the last syllable, the stretch of speech from the beginning of the last vowel to the end of the word. They found that by normalizing the rhyme by the speaking rate (calculated by comparing the duration of each phone with its mean) and accounting for differences in phones, they were able to make this feature more predictive.

In our work, we do not have enough data to normalize for the different phones on a speaker basis. Furthermore, normalizing using the entire corpus of speakers seems to do more damage than good. But we do normalize for speaker rate, using a moving window (Wightman and Ostendorf, 1991). The cumulative distribution curves are given in Figure 2.

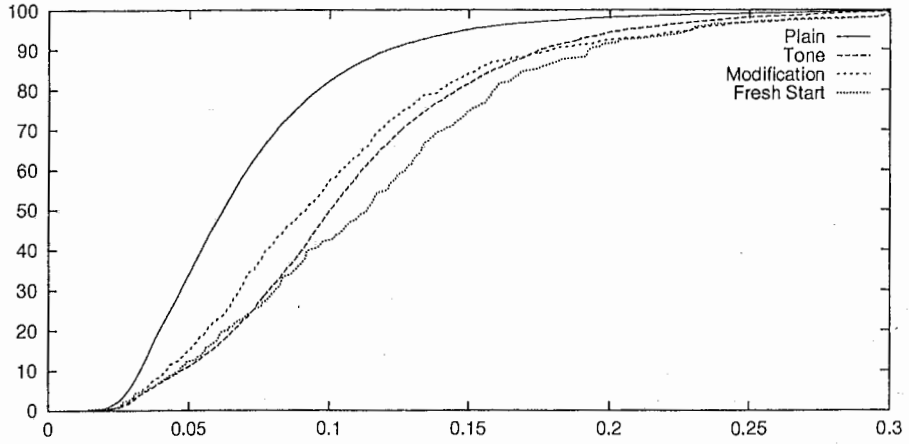

Figure 2: Rhyme duration normalized for speaking rate

The figure shows that the rhyme duration of plain transitions tends to be shorter than those for intonational phrase endings as Wightman *et al* found. However, it also shows that modification repairs and fresh starts tend to behave in the same manner as intonational phrases. In comparison to intonational phrase endings, fresh starts seem to have more rhyme lengthening and modification repairs less, with both repair types having a greater variance then intonational phrases do.

4.3 Pitch

One of the simplest measures is to take the ratio of the mean pitch of the preceding word over the mean pitch of the next word. This measure should also be robust to pitch tracking errors, variations due to individual phones, and speaker differences. The cumulative distribution of this measure is give in Figure 3. Ratios less than one indicate that the next word has a higher mean pitch than the previous word. The figure shows that fresh starts, modification

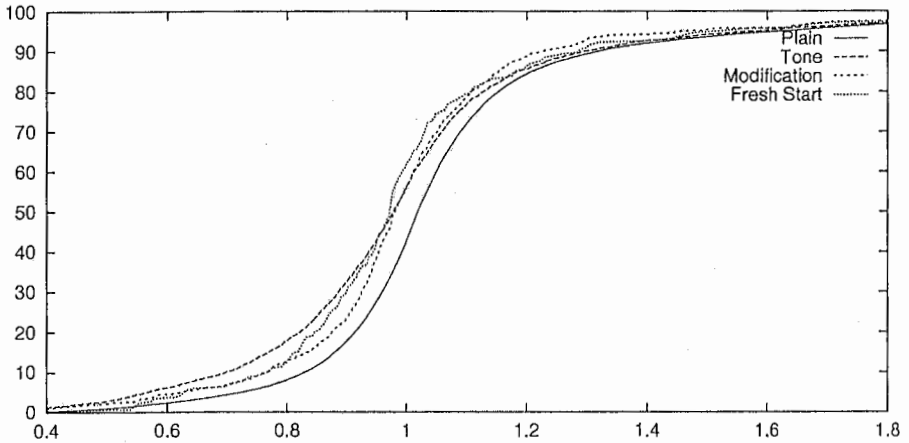

Figure 3: Ratio of mean pitch for previous word over next word.

repairs and intonational phrase boundaries tend to be accompanied by an increase in pitch

on the next word.

The shape of the pitch contour should be very predictive of intonational boundary tones. As well, many researchers have noted pitch contour effects with speech repairs. Following Wightman and Campbell (1994), we use the asymmetrical modal quadratic regression method developed by Hirst (Hirst and Espresser, 1993). This algorithm finds target points that define a spline fitting through the pitch contour. We use the target points to estimate the slope of the pitch contour at the end of the previous word and at the beginning of the next word.

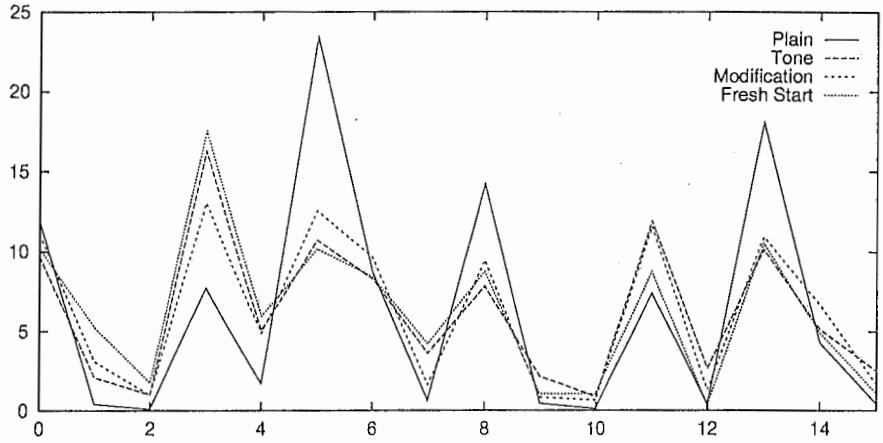

Figure 4: Peaks

Figure 4 gives the results of the analysis of slopes. The categories are arrived at by using the integers 0 for rising, 1 for falling, 2 for peak, and 3 for valley. We then multiplied the score of the next word by 4. At the scores of 5 and 13, there seems to be a preference for plain transitions. The score 5 is a fall for the previous word and a fall for the next word. So, where there is a continuing falling action on both words, we see that this more likely to be a plain transition.

The score of 13 corresponds to a fall for the previous word and a valley for the next word. Again we see that we have a continuing falling action, but part-way through the word, the contour starts to rise. This also is more indicative as the figure shows of a plain transition.

At the score of 3, we see a preference away from a plain transition. This score corresponds to a rise in the previous word followed by a valley in the next word.

4.4 Harmonic Ratio

The harmonic ratio is the difference in power between the first formant and the second format. For this work, we looked at the harmonic ratio of the last vowel in the previous word (Figure 5) and the first vowel in the next word (Figure 6). Here we are interested in whether there is any abrupt change. So, we normalize by subtracting off a running average of the harmonic ratios. This should also help us normalize for different speakers.

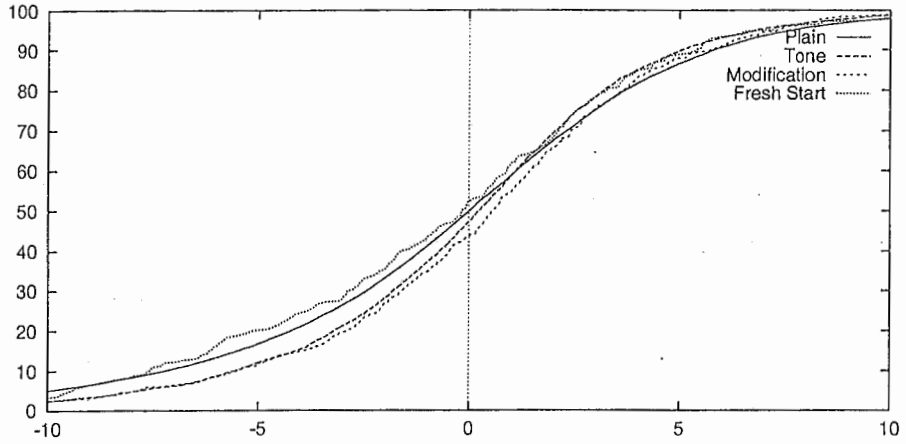

Figure 5: Harmonic ratio of previous word.

Figure 5 shows that fresh starts have a slightly lower harmonic ratio on the previous word, while modification repairs and intonational boundaries are slightly higher. The harmonic ratio of the next word (Figure 6 does not seem to be able to distinguish the transition types since all curves follow essentially the same curve.

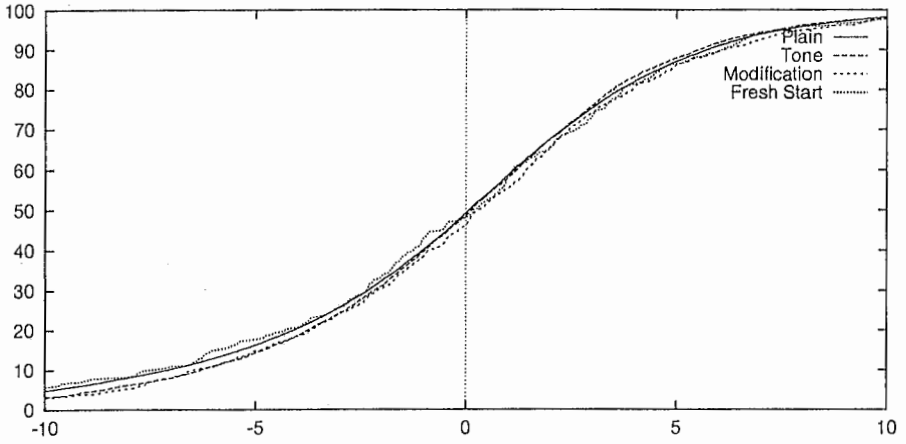

Figure 6: Harmonic ratio of next word.

Figure 7 shows the difference between the harmonic ratio of the last vowel of the previous word and the harmonic ratio of the first vowel of the next word. Here we see that fresh starts see to have more variance in their distribution than the other three types.

4.5 Spectral Tilt

Spectral tilt is the difference between the mid-band energy and the whole. In this study, we looked at the spectral tilt of the last vowel in the previous word (Figure 8) and the first vowel in the next word (Figure 9). These two features were normalized for the speaker's current average spectral tilt on vowels. We normalize by subtracting off a running average

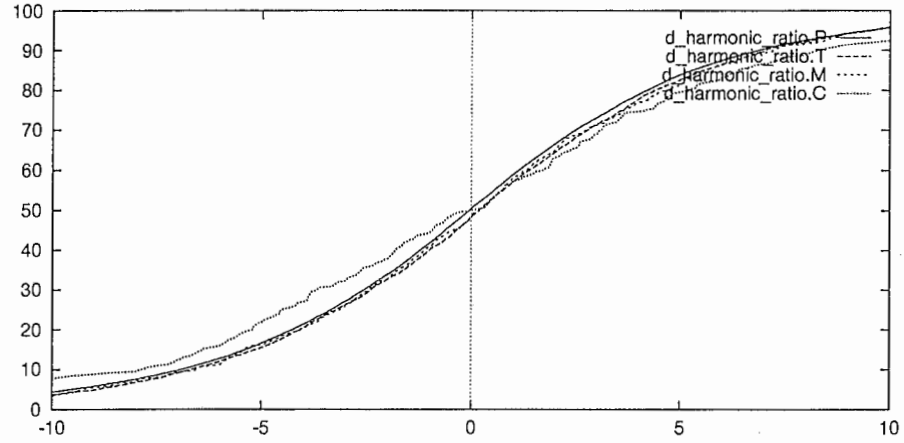

Figure 7: Difference in harmonic ratio between previous word and next word.

of the spectral tilts of the vowels. We also looked at the difference between the last vowel in the previous word and the first vowel in the next word (Figure 10).

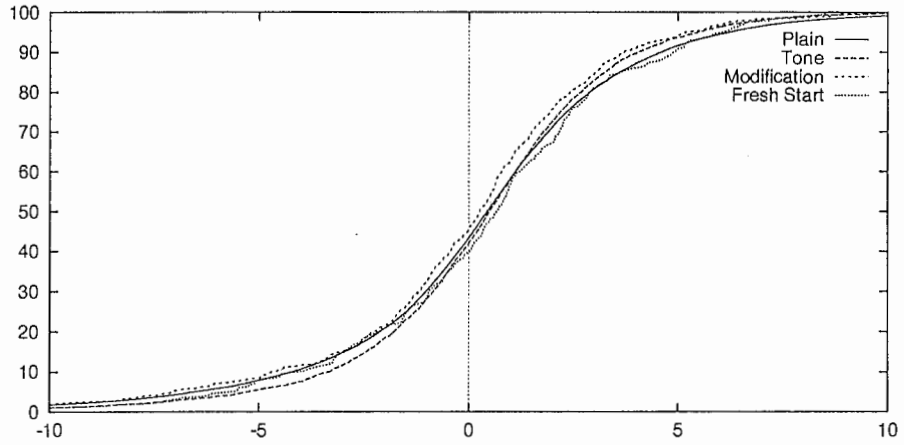

Figure 8: Spectral tilt of previous word.

In Figure 8, we give the spectral tilt on the last vowel preceding the transition type. Here we see that on the tone transitions do not have as much variance in their distribution in comparison to the plain transitions. It also seems that modification repairs have a lower spectral tilt than plain transitions.

In Figure 9, we give the spectral tilt on the first vowel after the transition type. Perhaps the most interesting of the distributions is the difference between the spectral tilt of the last vowel of the previous word and the first vowel of the next word (Figure 10). The most noticible aspect is that plain transitions have a much smaller variance than the other three distributions.

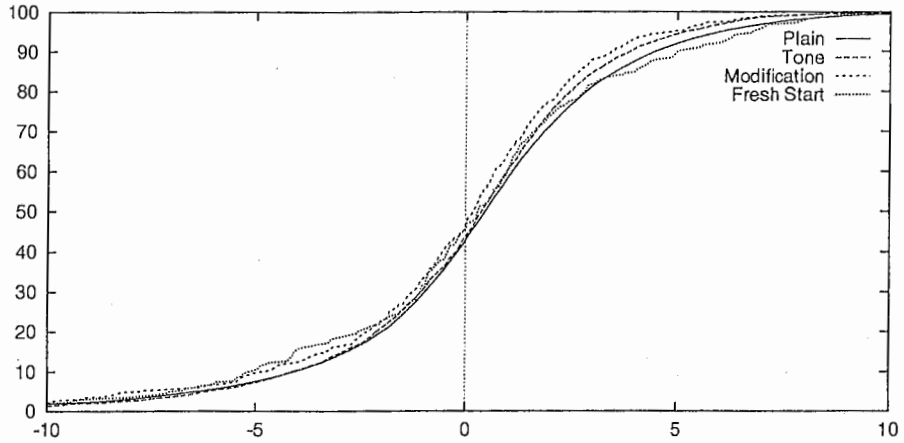

Figure 9: Spectral tilt of next word.

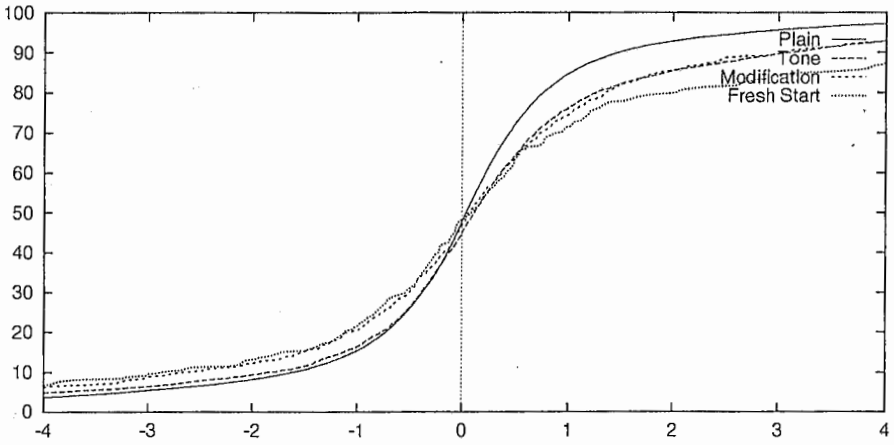

Figure 10: Difference between spectral tilt of previous word and next word.

4.6 Power

We looked at the maximum power in the last vowel of the previous word (Figure 11) and the first vowel in the next word (Figure 12). These features are normalized by the mean power over neighboring words. The distributions for the power on the last vowel (Figure 11) are rather surprising. The graph indicates that modification repairs and fresh starts tend to have more power on the final word than do fluent transitions, which is counter to intuition.

4.7 Lexical Stress

We also looked at whether there was lexical stress on the previous word and on the next word. Categories 2 and 3 indicate that the next word las lexical stress while categories 1 and 3 indicate that the previous word has lexical stress. As can be see in Figure 13, for intonational phrase endings, there is a tendency for the previous word to have lexical stress while the next word doesn't.

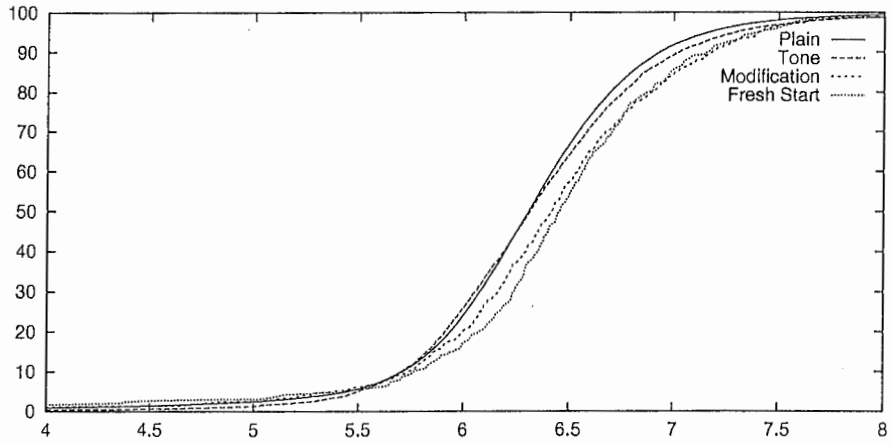

Figure 11: Maximum power in the last vowel of previous word.

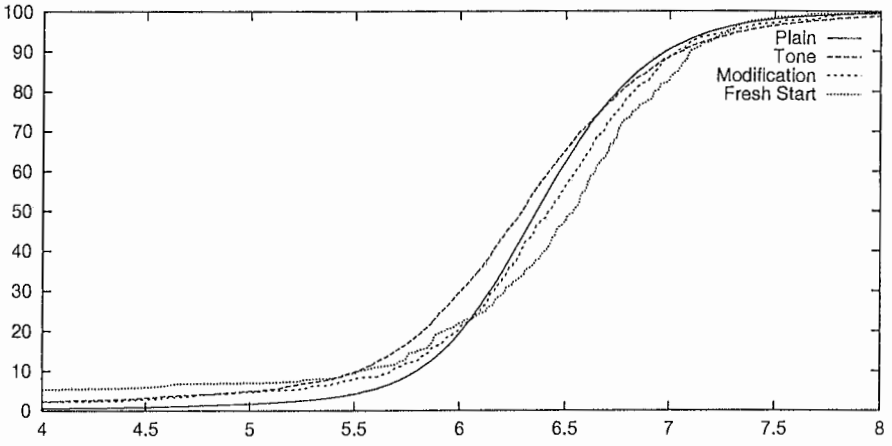

Figure 12: Maximum power in the first vowel of next word.

4.8 Remarks

We still need to do a lot more work in finding robust acoustic signals of speech repairs and intonational phrase boundaries. In particular, we need to find better ways to capture local disruptions in the acoustic signals, such as looking at the last syllable/vowel in the previous word and the first syllable/vowel in the next word. We also need to take care in normalizing these values. Speakers change the way they talk during a turn. What is important though are the points where their speech changes, and hence we either need to compare the feature across the transition point or with respect to a moving window. But if the window is too small, we will normalize out effect we are trying to study.

The above analysis might also be lacking in that it does not take into account the identify of the words. For instance, for a fresh start, the first word following the interruption point probably tends to be a function word, which would for instance have less power than a content word. So, in doing the power computations, it might be best to skip over function words.

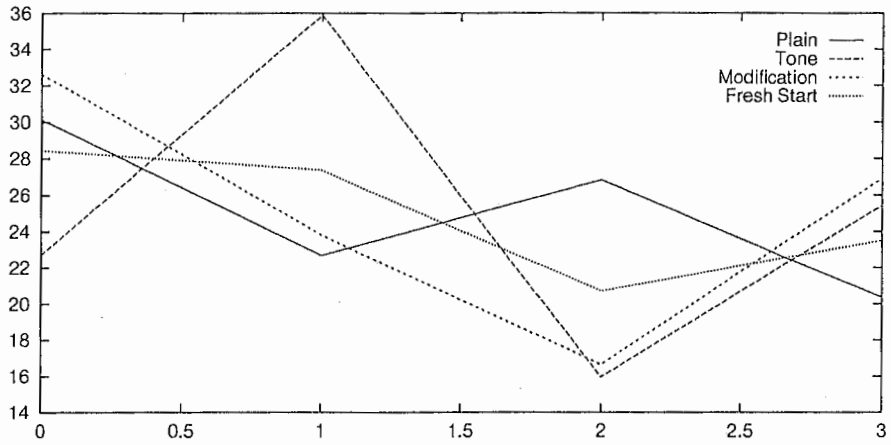

Figure 13: Lexical Stress

5 Combining the Clues

The acoustic clues that we mentioned are not independent. Hence to make use of these clues we need a technique that can combine without ignoring their interdependencies. So, we use CART (Breiman et al., 1984) to build decision trees. Table 1 gives the reduction in the tree impurity from using these clues, growing the tree to 80 leaves.

Clue	Reduction
Silence duration	0.04048
Rhyme duration of w_i normalized by Speaker Rate	0.00797
F0 slope analysis of w_i and w_{i+1}	0.00392
Normalized maximum power of first vowel of w_{i+1}	0.00239
Normalized maximum power of last vowel of w_i	0.00168
Ratio of mean pitch of w_i to mean pitch of w_{i+1}	0.00114
Presence of lexical stress on w_i and w_{i+1}	0.00091
Difference in spectral tilt of w_i and w_{i+1}	0.00074
Normalized harmonic ratio of w_i	0.00068
Difference in harmonic ratio of w_i and w_{i+1}	0.00064
Normalized spectral tilt of last vowel of w_i	0.00035
Normalized harmonic ratio of w_{i+1}	0.00019
Normalized spectral tilt of first vowel of w_{i+1}	0.00018

Table 1: Reduction in tree impurity for each acoustic clues.

6 Incorporating Acoustic Clues

In other work (Heeman and Allen, 1996), we present a model for detecting and correcting speech repairs and detecting intonational phrase boundaries using a statistical model based on a part-of-speech tagger. Modification repairs, fresh starts, and intonational phrase boundaries tend to have different category transition probabilities across the transition point than does fluent speech. By giving these distributions to a part-of-speech tagger, the tagger can decide the type of transition.

Part-of-speech tagging is the process of assigning to a word the category that is most probable given the sentential context (Church, 1988). The sentential context is typically approximated by only a set number of previous categories, usually one or two. Good part-of-speech results can be obtained using only the preceding category (Weischedel et al., 1993), which is what we will be using.

Figure 14 gives a simplified view of a Markov model for part-of-speech tagging, where C_i is a possible category for the ith word, w_i, and C_{i+1} is a possible category for word w_{i+1}. The category transition probability is simply the probability of category C_{i+1} following category C_i, which is written as $P(C_{i+1}|C_i)$. The probability of word w_{i+1} given category C_{i+1} is $P(w_{i+1}|C_{i+1})$. The category assignment that maximizes the product of these probabilities is taken to be the best category assignment.

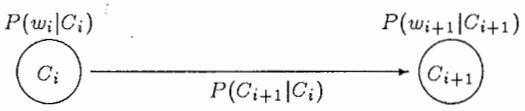

Figure 14: Markov Model of Part-of-Speech Tagging

Following Heeman and Allen, we make several independence assumptions about the occurrence of the transition types, which gives us the model depicted in Figure 15. This model allows us to view the problem as tagging null tokens between words with a transition type. Note that the context for category C_{i+1} is both C_i and T_i. So, T_i depends (indirectly) on the joint context of C_i and C_{i+1}, thus allowing the distributions of category co-occurrence to be used to model the transition type.

Modification repairs can be signaled by other indicators than just syntactic anomalies. For instance, the presence of word fragments and filled pauses, editing terms, silence duration and word matches also indicate their presence. This information can be added in by viewing the presence of these clues as part of the context to be used in computing the probabilities of the transition type. So, we replace $P(T_i|C_i)$ by $P(T_i|C_iF_iE_iM_i)$, where F_i indicates the presence of a word fragment, E_i indicates the presence of an editing term, and M_i indicates the presence of a word matching. If we make independence assumptions about the occurrence of these clues, we can rewrite this as the following.

$$P(T_i|C_i) \cdot P(T_i|F_i)/P(T_i) \cdot P(E_i|F_i)/P(T_i) \cdot \ldots$$

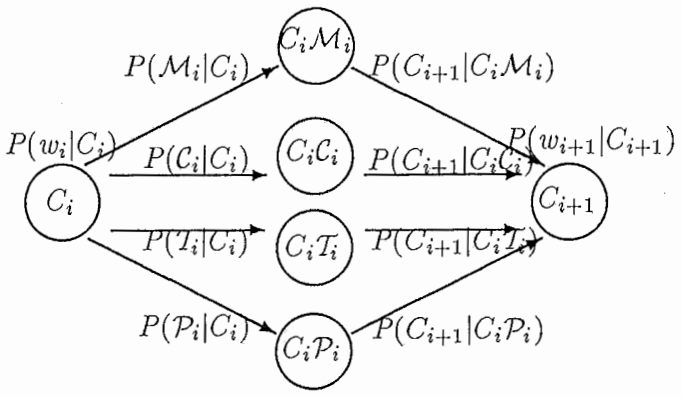

Figure 15: Statistical Model of Speech Repairs

To incorporate the acoustic clues discussed in the previous section, we simply view them as part of the context for deciding the transition type. By making independence assumptions about the co-occurrence of the acoustics clues A_i with the ones just discussed, we can incorporate the clues by multiplying in the following factor.

$$P(T_i|A_i)/P(T_i)$$

6.1 Results

Table 16 gives the results of a six-fold cross validation test. The corpus was divided into six parts, and for testing each part, the other five parts were used for training data. The first row gives the results of the tagging model without any acoustic clues. The second row gives the results of the model reported in (Heeman and Allen, 1996), which uses silence durations. The third gives the results from incorporating the full acoustic model (an earlier version than the one used for Table 1) generated by the CART analysis into the tagging model.

	Pos	Modification		Fresh Starts		Tones	
Model	Errors	Recall	Prec	Recall	Precision	Recall	Precision
No Acoustic Clues	2447	75.3	79.1	50.9	66.7	61.2	65.7
Baseline	2408	78.0	79.8	54.2	67.3	67.8	66.6
All Clues	2385	79.5	80.7	55.5	68.4	69.1	67.8

Figure 16: Comparison of performance of Tagging model using no acoustic clues, using silence duration, and using all of the acoustic clues.

As can be seen in the table, adding the full acoustic model to the tagging model improves the detecting of speech repairs and intonational phrase boundaries significantly. In comparison to the baseline, the model reported in (Heeman and Allen, 1996) that uses silence

durations, the improvement is less dramatic. This is because the pausal information gives the most information about the transition type of the acoustic clues that we have examined. Also, more work is needed in finding the best way to combine the clues. For instance, using the CART approach to combine rhyme durations and silence information gives worst results in the tagging model then simply modeling the silence duration with a probability density curve.

7 Comparison to Other Work

Wightman and Ostendorf (1994) have developed an acoustic model that labels intonational phrase endings as well as prominence. Developed and tested on speech read by a professional broadcaster, they achieved a recall rate of 78.1% and a precision of 76.8%. These figures are higher than our recall rate of intra-turn intonational phrase endings of 67.8% and precision of 66.6%. Differences between spontaneous dialogue and read speech account for some of the differences, while their detailed speaker-dependent acoustic modeling probably accounts for the rest.

Wang and Hirschberg (1992) also looked at detecting intonational phrase endings, running after syntactic analysis has been performed. Using automatically-labeled features, they were able to achieve a (cross-validated) success rate of 90%. Translated into recall and precision, they achieve a recall rate of $453/703 = 64.5\%$, and a precision of $453/512 = 88.5\%$. However, these results include 298 end of turns, in which only one was not a boundary. Excluding these, we arrive at a recall rate of 38.4% and a precision of 72.9%, in comparison to our recall rate of 67.8% and precision of 66.6%.

Nakatani and Hirschberg (1993) investigated using acoustic information to detect the interruption point of speech repairs. On a test set, they obtained a recall rate of 83.4% and a precision of 93.9% in detecting speech repairs using prosodic and lexical hand-transcribed annotations. However, all utterances in their training set and test set contained at least one repair, and repairs in their corpus occurred in about 5% of the utterances. So, it is unclear how this would scale up.

The SRI group (Bear, Dowding, and Shriberg, 1992) employed simple pattern matching techniques for detecting and correcting modification repairs. For detection, they were able to achieve a recall rate of 76%, and a precision of 62%, and they were able to find the correct repair 57% of the time, leading to an overall correction recall rate of 43% and correction precision of 50%. They also tried combining syntactic and semantic knowledge in a "parser-first" approach—first try to parse the input and if that fails, invoke repair strategies based on word patterns in the input. In a test set containing 26 repairs (Dowding et al., 1993), they obtained a detection recall rate of 42% and a precision of 84.6%; for correction, they obtained a correction recall rate of 30% and a recall rate of 62%.

8 Conclusion

In this paper we have shown that acoustic clues can be used to detect speech repairs and intonational phrase endings. Many of the same clues that give evidence for detecting intonational phrase endings are the same ones that can be used to detect speech repairs. Hence, for spontaneous speech, it is essential that both intonational phrase detection and speech repairs be modeled in by the same framework.

9 Acknowledgments

I wish to thank James Allen, Allen Black, Nick Campbell, Laurie Fais, Kyung-ho Loken-Kim, and Tsuyoshi Morimoto,

References

Allen, James F., Lenhart K. Schubert, George Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel Martin, Bradford Miller, Massimo Poesio, and David R. Traum. 1995. The Trains project: A case study in building a conversational planning agent. *Journal of Experimental and Theoretical AI*, 7:7–48. Also published as Trains TN 94-3 and TR 532, Computer Science Dept., U. Rochester, September 1994.

Bear, John, John Dowding, and Elizabeth Shriberg. 1992. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 56–63.

Bear, John, John Dowding, Elizabeth Shriberg, and Patti Price. 1993. A system for labeling self-repairs in speech. Technical Note 522, SRI International, February.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees.* Monterrey, CA: Wadsworth & Brooks.

Church, K. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pages 136–143, Febuary.

Clark, Herbert H., editor. 1992. *Arenas of Language Use.* University of Chicago Press and CSLI.

Clark, Herbert H. 1996. Using language. Unpublished manuscript.

Clark, Herbert H. and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39. Reprinted in (Clark, 1992), pages 107–143.

Dowding, John, Jean Mark Gawron, Doug Appelt, John Bear, Lynn Cherny, Robert Moore, and Douglas Moran. 1993. Gemini: A natural language system for spoken-language understanding. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, pages 54–61.

Entropic Research Laboratory, Inc., 1993. *WAVES+ Reference Manual.* Version 5.0.

Ford, Cecelia and Sandra Thompson. 1991. On projectability in conversation: Grammar, intonation, and semantics. Presented at the *Second International Cognitive Linguistics Association Conference*, August.

Heeman, Peter A. and James Allen. 1995. The Trains 93 dialogues. Trains Technical Note 94-2, Department of Computer Science, University of Rochester, March.

Heeman, Peter A. and James Allen. 1996. Annotating speech repairs. Unpublished manuscript.

Heeman, Peter A. and James F. Allen. 1995. The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium, April.

Heeman, Peter A. and James F. Allen. 1996. Detecting speech repairs and intonational phrase boundaries using a part-of-speech tagger. Unpublished manuscript.

Hirst, D. and R. Espresser. 1993. Automatic modeling of fundamental frequency using a quadratic spline function. *Travauxde l'Institue de Phonetique d'Aix*, 15:71–85.

Lickley, R. J. and E. G. Bard. 1992. Processing disfluent speech: Recognizing disfluency before lexical access. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pages 935–938, October.

Nakatani, Christine and Julia Hirschberg. 1993. A speech-first model for repair detection and correction. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, pages 46–53.

O'Shaughnessy, Douglas. 1992. Analysis of false starts in spontaneous speech. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pages 931–934, October.

Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, December.

Wang, Michelle Q. and Julia Hirschberg. 1992. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175–196.

Weischedel, Ralph, Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeff Palmucci. 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, 19(2):359–382.

Wightman, C. W. and M. Ostendorf. 1991. Automatic recognition of prosodic phrases. In *Proceddings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto.

Wightman, Colin W. and Nick Campbell. 1994. Automatic labeling of prosodic patterns. Technical Report TR-IT-0061, ATR Interpreting Telecommunications Research Laboratories, July.

Wightman, Colin W. and Mari Ostendorf. 1994. Automatic labeling of prosodic patterns. *IEEE Transactions on speech and audio processing*, October.

Wightman, Colin W., Stefanie Shattuck-Hufnagel, Mari Ostendorf, and Patti J. Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Accoustic Society of America*, 91(3):1707–1717, March.