A SPEECH-FIRST MODEL FOR REPAIR DETECTION AND CORRECTION

Christine Nakatani
Division of Applied Sciences
Harvard University
Cambridge, MA 02138
chn@das.harvard.edu

Julia Hirschberg
2D-450, AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974-0636
julia@research.att.com

Abstract

Interpreting fully natural speech is an important goal for spoken language understanding systems. However, while corpus studies have shown that about 10% of spontaneous utterances contain self-corrections, or REPAIRS, little is known about the extent to which cues in the speech signal may facilitate repair processing. We identify several cues based on acoustic and prosodic analysis of repairs in a corpus of spontaneous speech, and propose methods for exploiting these cues to detect and correct repairs. We test our acoustic-prosodic cues with other lexical cues to repair identification and find that precision rates of 89-93% and recall of 78-83% can be achieved, depending upon the cues employed, from a prosodically labeled corpus.

Introduction

Disfluencies in spontaneous speech pose serious problems for spoken language systems. First, a speaker may produce a partial word or FRAGMENT, a string of phonemes that does not form the complete intended word. Some fragments may coincidentally match words actually in the lexicon, such as *fly* in Example (1); others will be identified with the acoustically closest item(s) in the lexicon, as in Example (2).[1]

(1) What is the earliest **fli–** flight from Washington to Atlanta leaving on Wednesday September fourth?

(2) *Actual string*: What is the fare **fro–** on American Airlines fourteen forty three
Recognized string: With fare **four** American Airlines fourteen forty three

Even if all words in a disfluent segment are correctly recognized, failure to detect a disfluency may lead to interpretation errors during subsequent processing, as in Example (3).

(3) ... Delta leaving Boston seventeen twenty one arriving Fort Worth **twenty two** twenty one forty ...

Here, '*twenty two twenty one forty*' must be interpreted as a flight arrival time; the system must somehow choose among '21:40', '22:21', and '22:40'.

Although studies of large speech corpora have found that approximately 10% of spontaneous utterances contain disfluencies involving self-correction, or REPAIRS (Hindle, 1983; Shriberg et al., 1992), little is known about how to integrate repair processing with real-time speech recognition. In particular, the speech signal itself has been relatively unexplored as a source of processing cues for the detection and correction of repairs. In this paper, we present results from a study of the acoustic and prosodic characteristics of 334 repair utterances, containing 368 repair instances, from the ARPA Air Travel Information System (ATIS) database. Our results are interpreted within our "speech-first" framework for investigating repairs, the REPAIR INTERVAL MODEL (RIM). RIM builds upon Labov (1966) and Hindle (1983) by conceptually extending the EDIT SIGNAL HYPOTHESIS — that repairs are acoustically or phonetically marked at the point of interruption of fluent speech. After describing acoustic and prosodic characteristics of the repair instances in our corpus, we use these and other lexical cues to test the utility of our "speech-first" approach to repair identification on a prosodically labeled corpus.

Previous Computational Approaches

While self-correction has long been a topic of psycholinguistic study, computational work in this area has been sparse. Early work in computational linguistics treated repairs as one type of ill-formed input and proposed solutions based upon extensions to existing text parsing techniques such as augmented transition networks (ATNs), network-based semantic grammars, case frame grammars, pattern matching and deterministic parsers.

Recently, Shriberg et al. (1992) and Bear et al. (1992) have proposed a two-stage method for processing repairs. In the first stage, lexical pattern

[1] The presence of a word fragment in examples is indicated by the diacritic '–'. Self-corrected portions of the utterance appear in boldface. All examples in this paper are drawn from the ATIS corpus described below. Recognition output shown in Example (2) is from the system described in (Lee et al., 1990).

matching rules operating on orthographic transcriptions would be used to retrieve candidate repair utterances. In the second, syntactic, semantic, and acoustic information would filter true repairs from false positives found by the pattern matcher. Results of testing the first stage of this model, the lexical pattern matcher, are reported in (Bear et al., 1992): 309 of 406 utterance containing 'nontrivial' repairs in their 10,718 utterance corpus were correctly identified, while 191 fluent utterances were incorrectly identified as containing repairs. This represents recall of 76% with precision of 62%. Of the repairs correctly identified, the appropriate correction was found for 57%. Repair candidates were filtered and corrected by deleting a portion of the utterance based on the pattern matched, and then checking the syntactic and semantic acceptability of the corrected version using the syntactic and semantic components of the Gemini NLP system. Bear et al. (1992) also speculate that acoustic information might be used to filter out false positives for candidates matching two of their lexical patterns — repetitions of single words and cases of single inserted words — but do not report such experimentation.

This work promotes the important idea that automatic repair processing can be made more robust by integrating knowledge from multiple sources. Such integration is a desirable long-term goal. However, the working assumption that correct transcriptions will be available from speech recognizers is problematic, since current recognition systems rely primarily upon language models and lexicons derived from fluent speech to decide among competing acoustic hypotheses. These systems usually treat disfluencies in training and recognition as noise; moreover, they have no way of modeling word fragments, even though these occur in the majority of repairs. We term such approaches that rely on accurate transcription to identify repair candidates "text-first".

Text-first approaches have explored the potential contributions of lexical and grammatical information to automatic repair processing, but have largely left open the question of whether there exist acoustic and prosodic cues for repairs *in general*, rather than potential acoustic-prosodic filters for particular pattern subclasses. Our investigation of repairs addresses the problem of identifying such general acoustic-prosodic cues to repairs, and so we term our approach "speech-first". Finding such cues to repairs would provide early detection of repairs in recognition, permitting early pruning of the hypothesis space.

One proposal for repair processing that lends itself to both incremental processing and the integration of speech cues into repair detection is that of Hindle (1983), who defines a typology of repairs and associated correction strategies in terms of extensions to a deterministic parser. For Hindle, repairs can be (1) full sentence restarts, in which an entire utterance is reinitiated; (2) constituent repairs, in which one syntactic constituent (or part thereof) is replaced by another;[2] or (3) surface level repairs, in which identical strings appear adjacent to each other. An hypothesized acoustic-phonetic edit signal, "a markedly abrupt cut-off of the speech signal" (Hindle, 1983, p.123), is assumed to mark the interruption of fluent speech (cf. (Labov, 1966)). This signal is treated as a special lexical item in the parser input stream that triggers certain correction strategies depending on the parser configuration. Thus, in Hindle's system, repair detection is decoupled from repair correction, which requires only that the location of the interruption is stored in the parser state.

Importantly, Hindle's system allows for non-surface-based corrections and sequential application of correction rules (Hindle, 1983, p. 123). In contrast, simple surface deletion correction strategies cannot readily handle either repairs in which one syntactic constituent is replaced by an entirely different one, as in Example (4), or sequences of overlapping repairs, as in Example (5).

(4) I 'd like **to** a flight from Washington to Denver ...

(5) I 'd like to book a **reser– are there f–** is there a first class fare for the flight that departs at six forty p.m.

Hindle's methods achieved a success rate of 97% on a transcribed corpus of approximately 1,500 sentences in which the edit signal was orthographically represented and lexical and syntactic category assignments hand-corrected, indicating that, in theory, the edit signal can be computationally exploited for both repair detection and correction. Our "speech-first" investigation of repairs is aimed at determining the extent to which repair processing algorithms can rely on the edit signal hypothesis in practice.

The Repair Interval Model

To support our investigation of acoustic-prosodic cues to repair detection, we propose a "speech-first" model of repairs, the REPAIR INTERVAL MODEL (RIM). RIM divides the repair event into three consecutive temporal intervals and identifies time points within those intervals that are computationally critical. A full repair comprises three intervals, the REPARANDUM INTERVAL, the DISFLUENCY INTERVAL, and the REPAIR INTERVAL. Following Levelt (1983), we identify the REPARANDUM as the lexical material which is to be repaired. The end of the reparandum coincides with the termination of the fluent portion of the utterance, which we term the INTERRUPTION SITE (IS). The DISFLUENCY INTERVAL (DI) extends from the IS to the resumption of fluent speech, and may contain any combination of silence, pause fillers ('*uh*', '*um*'), or CUE PHRASES (e.g., '*Oops*'

[2]This is consistent with Levelt (1983)'s observation that the material to be replaced and the correcting material in a repair often share structural properties akin to those shared by coordinated constituents.

or '*I mean*'), which indicate the speaker's recognition of his/her performance error. The REPAIR INTERVAL corresponds to the utterance of the correcting material, which is intended to 'replace' the reparandum. It extends from the offset of the DI to the resumption of non-repair speech. In Example (6), for example, the reparandum occurs from 1 to 2, the DI from 2 to 3, and the repair interval from 3 to 4; the IS occurs at 2.

(6) Give me airlines **1** [**flying to Sa–**] **2** [SILENCE uh SILENCE] **3** [flying to Boston] **4** from San Francisco next summer that have business class.

RIM provides a framework for testing the extent to which cues from the speech signal contribute to the identification and correction of repair utterances. RIM incorporates two main assumptions of Hindle (1983): (1) correction strategies are linguistically rule-governed, and (2) linguistic cues must be available to signal when a disfluency has occurred and to 'trigger' correction strategies. As Hindle noted, if the processing of disfluencies were not rule-governed, it would be difficult to reconcile the infrequent intrusion of disfluencies on human speech comprehension, especially for language learners, with their frequent rate of occurrence in spontaneous speech. We view Hindle's results as evidence supporting (1). Our study tests (2) by exploring the acoustic and prosodic features of repairs that might serve as a form of edit signal for rule-governed correction strategies.

While Labov and Hindle proposed that an acoustic-phonetic cue might exist at precisely the IS, based on our analyses and on recent psycholinguistic experiments (Lickley et al., 1991), this proposal appears too limited. Crucially, in RIM, we extend the notion of edit signal to include any phenomenon which may contribute to the perception of an "abrupt cut-off" of the speech signal — including cues such as coarticulation phenomena, word fragments, interruption glottalization, pause, and prosodic cues which occur in the vicinity of the disfluency interval. RIM thus acknowledges the edit signal hypothesis, that some aspect of the speech signal may demarcate the computationally key juncture between the reparandum and repair intervals, while extending its possible acoustic and prosodic manifestations.

Acoustic-Prosodic Characteristics of Repairs

We studied the acoustic and prosodic correlates of repair events as defined in the RIM framework with the aim of identifying potential cues for automatic repair processing, extending a pilot study reported in (Nakatani and Hirschberg, 1993). Our corpus for the current study consisted of 6,414 utterances produced by 123 speakers from the ARPA Airline Travel and Information System (ATIS) database (MADCOW, 1992) collected at AT&T, BBN, CMU, SRI, and TI. 334 (5.2%) of these utterances contain at least one repair, where repair is defined as the self-correction of one or more phonemes (up to and including sequences of words) in an utterance.[3] Orthographic transcriptions of the utterances were prepared by ARPA contractors according to standardized conventions. The utterances were labeled at Bell Laboratories for word boundaries and intonational prominences and phrasing following Pierrehumbert's description of English intonation (Pierrehumbert, 1980). Also, each of the three RIM intervals and prosodic and acoustic events within those intervals were labeled.

Identifying the Reparandum Interval

Our acoustic and prosodic analysis of the reparandum interval focuses on acoustic-phonetic properties of word fragments, as well as additional phonetic cues marking the reparandum offset. From the point of view of repair detection and correction, acoustic-prosodic cues to the onset of the reparandum would clearly be useful in the choice of appropriate correction strategy. However, recent perceptual experiments indicate that humans do not detect an oncoming disfluency as early as the onset of the reparandum (Lickley et al., 1991; Lickley and Bard, 1992). Subjects *were* generally able to detect disfluencies before lexical access of the first word in the repair. However, since only a small number of the test stimuli employed in these experiments contained reparanda ending in word fragments (Lickley et al., 1991), it is not clear how to generalize results to such repairs. In our corpus, 74% of all reparanda end in word fragments.[4]

Since the majority of our repairs involve word fragmentation, we analyzed several lexical and acoustic-phonetic properties of fragments for potential use in fragment identification. Table 1 shows the broad word class of the speaker's intended word for each fragment, where the intended word was recoverable. There is

Lexical Class	Tokens	%
Content	121	42%
Function	12	4%
Untranscribed	155	54%

Table 1: Lexical Class of Word Fragments at Reparandum Offset (N=288)

a clear tendency for fragmentation at the reparandum offset to occur in content words rather than function words.

[3] In our pilot study of the SRI and TI utterances only, we found that repairs occurred in 9.1% of utterances (Nakatani and Hirschberg, 1993). This rate is probably more accurate than the 5.2% we find in our current corpus, since repairs for the pilot study were identified from more detailed transcriptions than were available for the larger corpus.

[4] Shriberg et al. (1992) found that 60.2% of repairs in their corpus contained fragments.

Table 2 shows the distribution of fragment repairs by length. 91% of fragments in our corpus are one syllable or less in length. Table 3 shows the distribution of initial phonemes for all words in the corpus of 6,414 ATIS sentences, and for all fragments, single syllable fragments, and single consonant fragments in repair utterances. From Table 3 we see that single consonant fragments occur more than six times as often as fricatives than as stops. However, fricatives and stops occur almost equally as the initial consonant in single syllable fragments. Furthermore, we observe two divergences from the underlying distributions of initial phonemes for all words in the corpus. Vowel-initial words show less tendency and fricative-initial words show a greater tendency to occur as fragments, relative to the underlying distributions for those classes.

Syllables	Tokens	%
0	113	39%
1	149	52%
2	25	9%
3	1	0.3%

Table 2: Length of Reparandum Offset Word Fragments (N=288)

Class	% of Words	% of Frags	% of One Syll Frags	% of One Cons Frags
stop	23%	23%	30%	11%
vowel	25%	13%	19%	0%
fric	33%	45%	28%	73%
nasal/ glide/ liquid	18%	17%	20%	15%
h	1%	2%	4%	1%
N	64896	288	148	114

Table 3: Feature Class of Initial Phoneme in Fragments by Fragment Length

Two additional acoustic-phonetic cues, glottalization and coarticulation, may help in fragment identification. Bear et al. (1992) note that INTERRUPTION GLOTTALIZATION (irregular glottal pulses) sometimes occurs at the reparandum offset. This form of glottalization is acoustically distinct from LARYNGEALIZATION (creaky voice), which often occurs at the end of prosodic phrases; GLOTTAL STOPS, which often precede vowel-initial words; and EPENTHETIC GLOTTALIZATION. In our corpus, 30.2% of reparanda offsets are marked by interruption glottalization.[5] Although interruption glottalization is usually associated with fragments, not all fragments are glottalized. In our database, 62% of fragments are *not* glottalized, and 9% of glottalized reparanda offsets are *not* fragments.

[5]Shriberg et al. (1992) report glottalization on 24 of 25 vowel-final fragments.

Also, sonorant endings of fragments in our corpus sometimes exhibit coarticulatory effects of an unrealized subsequent phoneme. When these effects occur with a following pause (see below), they can be used to distinguish fragments from full phrase-final words — such as *'fli–'* from *'fly'* in Example (1).

To summarize, our corpus shows that most reparanda offsets end in word fragments. These fragments are usually fragments of content words (based upon transcribers' identification of intended words in our corpus), are rarely more than one syllable long, exhibit different distributions of initial phoneme class depending on their length, and are sometimes glottalized and sometimes exhibit coarticulatory effects of missing subsequent phonemes. These findings suggest that it is unlikely that word-based recognition models can be applied directly to the problem of fragment identification. Rather, models for fragment identification might make use of initial phoneme distributions, in combination with information on fragment length and acoustic-phonetic events at the IS. Inquiry into the articulatory bases of several of these properties of self-interrupted speech, such as glottalization and initial phoneme distributions, may further improve the modeling of fragments.

Identifying the Disfluency Interval

In the RIM model, the DI includes all cue phrases and filled and unfilled pauses from the offset of the reparandum to the onset of the repair. The literature contains a number of hypotheses about this interval (cf. (Blackmer and Mitton, 1991). For our corpus, pause fillers or cue words, which have been hypothesized as repair cues, occur within the DI for only 9.8% (332/368) of repairs, and so cannot be relied on for repair detection. Our findings do, however, support a new hypothesis associating fragment repairs and the duration of pause following the IS.

Table 4 shows the average duration of 'silent DI's (those not containing pause fillers or cue words) compared to that of fluent utterance-internal silent pauses for the TI utterances. Overall, silent DIs are shorter than fluent pauses ($p<.001$, tstat=4.60, df=1516). If we analyze repair utterances based on occurrence of fragments, the DI duration for fragment repairs is significantly shorter than for nonfragments ($p<.001$, tstat=3.36, df=330). The fragment repair DI duration is also significantly shorter than fluent pause intervals

Pausal Juncture	Mean	Std Dev	N
Fluent	513 msec	676 msec	1186
DI	333 msec	417 msec	332
Frags	292 msec	379 msec	255
Non-frags	471 msec	502 msec	77

Table 4: Duration of Silent DIs vs. Utterance-Internal Fluent Pauses

($p<.001$, tstat=5.05, df=1439), while there is no significant difference between nonfragment DIs and fluent utterances. So, DIs in general appear to be distinct from fluent pauses, and the duration of DIs in fragment repairs might also be exploited to identify these cases as repairs, as well as to distinguish them from nonfragment repairs. Thus, pausal duration may serve as a general acoustic cue for repair detection, particularly for the class of fragment repairs.

Identifying the Repair

Several influential studies of acoustic-prosodic repair cues have relied upon lexical, semantic, and pragmatic definitions of repair types (Levelt and Cutler, 1983; Levelt, 1983). Levelt & Cutler (1983) claim that repairs of erroneous information (ERROR REPAIRS) are marked by increased intonational prominence on the correcting information, while other kinds of repairs, such as additions to descriptions (APPROPRIATENESS REPAIRS), generally are not. We investigated whether the repair interval is marked by special intonational prominence relative to the reparandum for all repairs in our corpus and for these particular classes of repair.

To obtain objective measures of relative prominence, we compared absolute f0 and energy in the sonorant center of the last accented lexical item in the reparandum with that of the first accented item in the repair interval.[6] We found a small but reliable increase in f0 from the end of the reparandum to the beginning of the repair (mean=4.1 Hz, $p<.01$, tstat=2.49, df=327). There was also a small but reliable increase in amplitude across the DI (mean=+1.5 db, $p<.001$, tstat=6.07, df=327). We analyzed the same phenomena across utterance-internal fluent pauses for the ATIS TI set and found no reliable differences in either f0 or intensity, although this may have been due to the greater variability in the fluent population. And when we compared the f0 and amplitude changes from reparandum to repair with those observed for fluent pauses, we found no significant differences between the two populations.

So, while differences in f0 and amplitude exist between the reparandum offset and the repair onset, we conclude that these differences are too small help distinguish repairs from fluent speech. Although it is not entirely straightforward to compare our objective measures of intonational prominence with Levelt and Cutler's perceptual findings, our results provide only weak support for theirs. And while we find small but significant changes in two correlates of intonational prominence, the distributions of change in f0 and energy for our data are unimodal; when we further test subclasses of Levelt and Cutler's error repairs and appropriateness repairs, statistical analysis does *not* support Levelt and Cutler's claim that the former — and only the former — group is intonationally 'marked'.

Previous studies of disfluency have paid considerable attention to the vicinity of the DI but little to the repair offset. Although we did not find comparative intonational prominence across the DI to be a promising cue for repair *detection*, our RIM analysis uncovered one general intonational cue that may be of use for repair *correction*, namely the prosodic phrasing of the repair interval. We propose that phrase boundaries at the repair offset can serve to delimit the region over which subsequent correction strategies may operate.

We tested the idea that repair interval offsets are intonationally marked by either minor or major prosodic phrase boundaries in two ways. First, we used the phrase prediction procedure reported by Wang & Hirschberg (1992) to estimate whether the phrasing at the repair offset was predictable according to a model of fluent phrasing.[7] Second, we analyzed the syntactic and lexical properties of the first major or minor intonational phrase including all or part of the repair interval to determine whether such phrasal units corresponded to different types of repairs in terms of Hindle's typology.

The first analysis tested the hypothesis that repair interval offsets are intonationally delimited by minor or major prosodic phrase boundaries. We found that the repair offset co-occurs with minor phrase boundaries for 49% of repairs in the TI set. To see whether these boundaries were distinct from those in fluent speech, we compared the phrasing of repair utterances with the phrasing predicted for the corresponding corrected version of the utterance identified by ATIS transcribers. For 40% of all repairs, an observed boundary occurs at the repair offset where one is predicted; and for 33% of all repairs, no boundary is observed where none is predicted. For the remaining 27% of repairs for which predicted phrasing diverged from observed, in 10% of cases a boundary occurred where none was predicted and in 17%, no boundary occurred when one was predicted.

In addition to differences at the repair offset, we also found more general differences from predicted phrasing over the entire repair interval, which we hypothesize may be partly understood as follows: Two strong predictors of prosodic phrasing in fluent speech are syntactic constituency (Cooper and Sorenson, 1977; Gee and Grosjean, 1983; Selkirk, 1984), especially the relative inviolability of noun phrases (Wang and Hirschberg, 1992), and the length of prosodic phrases (Gee and Grosjean, 1983; Bachenko

[6] We performed the same analysis for the last and first syllables in the reparandum and repair, respectively, and for normalized f0 and energy; results did not substantially differ from those presented here.

[7] Wang & Hirschberg use statistical modeling techniques to predict phrasing from a large corpus of labeled ATIS speech; we used a prediction tree that achieves 88.4% accuracy on the ATIS TI corpus using only features whose values could be calculated via automatic text analysis. Results reported here are for prediction on only TI repair utterances.

and Fitzpatrick, 1990). On the one hand, we found occurrences of phrase boundaries at repair offsets which occurred within larger NPs, as in Example (7), where it is precisely the noun modifier — not the entire noun phrase — which is corrected.[8]

(7) Show me all **n–** | round-trip flights | from Pittsburgh | to Atlanta.

We speculate that, by marking off the modifier intonationally, a speaker may signal that operations relating just this phrase to earlier portions of the utterance can achieve the proper correction of the disfluency. We also found cases of 'lengthened' intonational phrases in repair intervals, as illustrated in the single-phrase reparandum in (8), where the corresponding fluent version of the reparandum is predicted to contain four phrases.

(8) **What airport is it | is located** | what is the name of the airport located in San Francisco

Again, we hypothesize that the role played by this unusually long phrase is the same as that of early phrase boundaries in NPs discussed above. In both cases, the phrase boundary delimits a meaningful unit for subsequent correction strategies. For example, we might understand the multiple repairs in (8) as follows: First the speaker attempts a VP repair, with the repair phrase delimited by a single prosodic phrase '*is located*'. Then the initially repaired utterance '*What airport is located*' is itself repaired, with the reparadum again delimited by a single prosodic phrase, '*What is the name of the airport located in San Francisco*'.

In the second analysis of lexical and syntactic properties, we found three major classes of phrasing behaviors, all involving the location of the first phrase boundary after the repair onset: First, for 44% (163/368) of repairs, the repair offset we had initially identified[9] coincides with a phrase boundary, which can thus be said to mark off the repair interval. Of the remaining 205 repairs, more than two-thirds (140/205) have the first phrase boundary after the repair onset at the right edge of a syntactic constituent. We propose that this class of repairs should be identified as constituent repairs, rather than the lexical repairs we had initially hypothesized. For the majority of these constituent repairs (79%, 110/140), the repair interval contains a well-formed syntactic constituent (see Table 5). If the repair interval does *not* form a syntactic constituent, it is most often an NP-internal repair (77%, 23/30). The third class of repairs includes those in which the first boundary after the repair onset occurs neither at the repair offset nor at the right edge of a syntactic constituent. This class contains surface or lexical

[8]Prosodic boundaries in examples are indicated by '|'.

[9]Note crucially here that, in labeling repairs which might be viewed as either constituent or lexical, we preferred the shorter lexical analysis by default.

Repair Constituent	Tokens	%
Sentence	24	22%
Verb phrase	7	6%
Participial phrase	6	5%
Noun phrase	38	35%
Prepositional phrase	34	31%
Relative clause	1	0.9%

Table 5: Distribution of Syntactic Categories for Constituent Repairs (N=110)

repairs (where the first phrase boundary in the repair interval delimits a sequence of one or more repeated words), phonetic errors, word insertions, and syntactic reformulations (as in Example (4)). It might be noted here that, in general, repairs involving correction of either verb phrases or verbs are far less common than those involving noun phrases, prepositional phrases, or sentences.

We briefly note evidence against one alternative (although not mutually exclusive) hypothesis, that the region to be delimited correction strategies is marked not by a phrase boundary near the repair offset, but by a phrase boundary at the onset of the reparandum. In other words, it may be the reparandum interval, not the repair interval, that is intonationally delimited. However, it is often the case that the last phrase boundary before the IS occurs at the left edge of a major syntactic constituent (42%, (87/205), even though major constituent repairs are about one third as frequent in this corpus (15%, 31/205). In contrast, phrase boundaries occur at the left edge of minor constituents 27% (55/205) of the time, whereas minor constituent repairs make up 39% (79/205) of the subcorpus at hand. We take these figures as general evidence against the outlined alternative hypothesis, establishing that the demarcation repair offset is a more productive goal for repair processing algorithms.

Investigation of repair phrasing in other corpora covering a wider variety of genres is needed in order to assess the generality of these findings. For example, 35% (8/23) of NP-internal constituent repairs occurred within cardinal compounds, which are prevalent in the ATIS corpus due to its domain. The preponderance of temporal and locative prepositional phrases may also be attributed to the nature of the task and domain. Nonetheless, the fact that repair offsets in our corpus are marked by intonational phrase boundaries in such a large percentage of cases (82.3%, 303/368), suggests that this is a possibility worth pursuing.

Predicting Repairs from Acoustic and Prosodic Cues

Despite the small size of our sample and the possibly limited generality of our corpus, we were interested to see how well the characterization of repairs derived

from RIM analysis of the ATIS corpus would transfer to a predictive model for repairs in that domain. We examined 374 ATIS repair utterances, including the 334 upon which the descriptive study presented above was based. We used the 172 TI and SRI repair utterances from our earlier pilot study (Nakatani and Hirschberg, 1993) as training date; these served a similar purpose in the descriptive analysis presented above. We then tested on the additional 202 repair utterances, which contained 223 repair instances. In our predictions we attemped to distinguish repair IS from fluent phrase boundaries (collapsing major and minor boundaries), non-repair disfluencies,[10] and simple word boundaries. We considered every word boundary to be a potential repair site.[11] Data points are represented below as ordered pairs $<w_i,w_j>$, where w_i represents the lexical item to the left of the potential IS and w_j represents that on the right.

For each $<w_i,w_j>$, we examined the following features as potential IS predictors: (a) duration of pause between w_i and w_j; (b) occurrence of a word fragment(s) within $<w_i,w_j>$; (c) occurrence of a filled pause in $<w_i,w_j>$; (d) amplitude (energy) peak within w_i, both absolute and normalized for the utterance; (e) amplitude of w_i relative to w_{i-1} and to w_j; (f) absolute and normalized f0 of w_i; (g) f0 of w_i relative to w_{i-1} and to w_j; and (h) whether or not w_i was accented, deaccented, or deaccented and cliticized. We also simulated some simple pattern matching strategies, to try to determine how acoustic-prosodic cues might interact with lexical cues in repair identification. To this end, we looked at (i) the distance in words of w_i from the beginning and end of the utterance; (j) the total number of words in the utterance; and (k) whether w_i or w_{i-1} recurred in the utterance within a window of three words after w_i. We were unable to test all the acoustic-prosodic features we examined in our descriptive analysis, since features such as glottalization and coarticulatory effects had not been labeled in our data base for locations other than DIS. Also, we used fairly crude measures to approximate features such as change in f0 and amplitude, since these too had been precisely labeled in our corpus only for repair locations and not for fluent speech.[12]

We trained prediction trees, using Classification and Regression Tree (CART) techniques (Brieman et al., 1984), on our 172-utterance training set. We first included all our potential identifiers as possible predictors. The resulting (automatically generated) decision tree was then used to predict IS locations in our 202-utterance test set. This procedure identified 186 of the 223 repairs correctly, while predicting 12 false positives and omitting 37 true repairs, for a recall of 83.4% and precision of 93.9%. Fully 177 of the correctly identified ISs were identified via presence of word fragments as well as duration of pause in the DI. Repairs not containing fragments were identified from lexical matching plus pausal duration in the DI.

Since the automatic identification of word fragments from speech is an unsolved problem, we next omitted the fragment feature and tried the prediction again. The best prediction tree, tested on the same 202-utterance test set, succeeded in identifying 174 of repairs correctly — in the absence of fragment information — with 21 false positives and 49 omissions (78.1% recall, 89.2% precision). The correctly identified repairs were all characterized by constraints on duration of pause in the DI. Some were further identified via presence of lexical match to the right of w_i within the window of three described above, and word position within utterance. Those repairs in which no lexical match was identified were characterized by lower amplitude of w_i relative to w_j and cliticization or deaccenting of w_i. Still other repairs were characterized by more complex series of lexical and acoustic-prosodic constraints.

These results are, of course, very preliminary. Larger corpora must certainly be examined and more sophisticated versions of the crude measures we have used should be employed. However, as a first approximation to the characterization of repairs via both acoustic-prosodic and lexical cues, we find these results encouraging. In particular, our ability to identify repair sites successfully without relying upon the identification of fragments as such seems promising, although our analysis of fragments suggests that there may indeed be ways of identifying fragment repairs, via their relatively short DI, for example. Also, the combination of general acoustic-prosodic constraints with lexical pattern matching techniques as a strategy for repair identification appears to gain some support from our predictions. Further work on prediction modeling may suggest ways of combining these lexical and acoustic-prosodic cues for repair processing.

Discussion

In this paper, we have presented a "speech-first" model, the Repair Interval Model, for studying repairs in spontaneous speech. This model divides the repair event into a reparandum interval, a disfluency interval, and a repair interval. We have presented empirical results from acoustic-phonetic and prosodic analysis of a corpus of repairs in spontaneous speech, indicating that reparanda offsets end in word fragments, usually of (intended) content words, and that these fragments tend to be quite short and to exhibit particular acoustic-phonetic characteristics. We found that the disfluency

[10] These had been marked independently of our study and including all events with some phonetic indicator of disfluency which was not involved in a self-repair, such as hesitations marked with audible breath or sharp cut-off.

[11] We also included utterance-final boundaries as data points.

[12] We used uniform measures for prediction, however, for both repair sites and fluent regions.

interval can be distinguished from intonational phrase boundaries in fluent speech in terms of duration of pause, and that fragment and nonfragment repairs can also be distinguished from one another in terms of the duration of the disfluency interval. For our corpus, repair onsets can be distinguished from reparandum offsets by small but reliable differences in f0 and amplitude, and repair intervals differ from fluent speech in their characteristic prosodic phrasing. We tested our results by developing predictive models for repairs in the ATIS domain, using CART analysis; the best performing prediction strategies, trained on a subset of our data, identified repairs in the remaining utterances with recall of 78-83% and precision of 89-93%, depending upon features examined.

Acknowledgments

We thank John Bear, Barbara Grosz, Don Hindle, Chin Hui Lee, Robin Lickley, Andrej Ljolje, Jan van Santen, Stuart Shieber, and Liz Shriberg for advice and useful comments. CART analysis employed software written by Daryl Pregibon and Michael Riley. Speech analysis was done with Entropic Research Laboratory's WAVES software.

REFERENCES

J. Bachenko and E. Fitzpatrick. 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3):155–170.

John Bear, 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*, pages 56–63, Newark DE. Association for Computational Linguistics.

Elizabeth R. Blackmer and Janet L. Mitton. 1991. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39:173–194.

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

W. E. Cooper and J. M. Sorenson. 1977. Fundamental frequency contours at syntactic boundaries. *Journal of the Acoustical Society of America*, 62(3):683–692, September.

J. P. Gee and F. Grosjean. 1983. Performance structure: A psycholinguistic and linguistic apprasial. *Cognitive Psychology*, 15:411–458.

Donald Hindle. 1983. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting*, pages 123–128, Cambridge MA. Association for Computational Linguistics.

William Labov. 1966. On the grammaticality of everyday speech. Paper Presented at the Linguistic Society of America Annual Meeting.

C.-H. Lee, L. R. Rabiner, R. Pieraccini, and J. Wilpon. 1990. Acoustic modeling for large vocabulary speech recognition. *Computer Speech and Language*, 4:127–165, April.

William Levelt and Anne Cutler. 1983. Prosodic marking in speech repair. *Journal of Semantics*, 2:205–217.

William Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14:41–104.

R. J. Lickley and E. G. Bard. 1992. Processing disfluent speech: Recognising disfluency before lexical access. In *Proceedings of the International Conference on Spoken Language Processing*, pages 935–938, Banff, October. ICSLP.

R. J. Lickley, R. C. Shillcock, and E. G. Bard. 1991. Processing disfluent speech: How and when are disfluencies found? In *Proceedings of the Second European Conference on Speech Communication and Technology, Vol. III*, pages 1499–1502, Genova, September. Eurospeech-91.

MADCOW. 1992. Multi-site data collection for a spoken language corpus. In *Proceedings of the Speech and Natural Language Workshop*, pages 7–14, Harriman NY, February. DARPA, Morgan Kaufmann.

Christine Nakatani and Julia Hirschberg. 1993. A speech-first model for repair identification in spoken language systems. In *Proceedings of the ARPA Workshop on Human Language Technology*, Plainsboro, March. ARPA.

Janet B. Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation.* Ph.D. thesis, Massachusetts Institute of Technology, September. Distributed by the Indiana University Linguistics Club.

E. O. Selkirk. 1984. Phonology and syntax: The relation between sound and structure. In T. Freyjeim, editor, *Nordic Prosody II: Proceedings of the Second Symposium on Prosody in the Nordic language*, pages 111–140, Trondheim. TAPIR.

Elizabeth Shriberg, John Bear, and John Dowding. 1992. Automatic detection and correction of repairs in human-computer dialog. In *Proceedings of the Speech and Natural Language Workshop*, pages 419–424, Harriman NY. DARPA, Morgan Kaufmann.

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